ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 35 (2002) 142–150

Journal of Biomedical Informatics

www.academicpress.com

Methodological Review

Comparative genomics approaches to study organism similarities and differences

Liping Wei,[a,b,*] Yueyi Liu,[c] Inna Dubchak,[d] John Shon,[a] and John Park[a]

[a] *Nexus Genomics, Inc., 229 Polaris Ave., Suite 6, Mountain View, CA 94043, USA*
[b] *Hang Zhou Center, Beijing Genomics Institute, Hangzhou, China*
[c] *Stanford Medical Informatics, 251 Campus Dr. X215, Stanford University, Stanford CA 94305, USA*
[d] *Genome Sciences Department, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

Received 7 June 2002

Abstract

Comparative genomics is a large-scale, holistic approach that compares two or more genomes to discover the similarities and differences between the genomes and to study the biology of the individual genomes. Comparative studies can be performed at different levels of the genomes to obtain multiple perspectives about the organisms. We discuss in detail the type of analyses that offer significant biological insights in the comparisons of (1) genome structure including overall genome statistics, repeats, genome rearrangement at both DNA and gene level, synteny, and breakpoints; (2) coding regions including gene content, protein content, orthologs, and paralogs; and (3) noncoding regions including the prediction of regulatory elements. We also briefly review the currently available computational tools in comparative genomics such as algorithms for genome-scale sequence alignment, gene identification, and nonhomology-based function prediction.

1. Introduction

Biomedicine today has a powerful new resource for discovery—a rapidly growing number of sequenced genomes. It was only as recent as 1995 when the first complete genome sequence of a free-living organism—the bacterium *Haemophilus influenzae Rd*—was published [1]. Since then, through the large-scale DNA sequencing efforts of many public and private organizations, the complete genomes of 15 archaea, 67 bacteria, and 8 eukaryota have been revealed (as of July 2002) [2]. In addition, the draft genome sequences of several major organisms have become available, in particular, the draft genome sequences of human and rice [3–6]. The sequencing of the genomes of about 800 other organisms is currently in progress. Table 1 compares the sizes of several complete and draft sequences of archaea, bacteria, and eukaryotae genomes, and showcases both the large size of some of the eukaryotae genomes and the vast range of genome sizes. A full list of genomes, completed or in progress, and their sequences can be retrieved from the NCBI genome resources at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome.

There is undoubtedly much excitement over the sequencing of a complete genome. It is argued, however, that a genome taken in isolation from a single organism does not reveal much by itself. Genomes, and genes, need to be studied in comparison with other species (or subspecies, or strains), in the phylogenetic context of the evolutionary process [7]. Comparative genomics is a large-scale, holistic approach that compares two or more genomes to discover the similarities and differences between the genomes and to study the biology of the individual genomes. Comparative genomics applies to whole genomes or syntenic regions of different species, different subspecies, or different strains of the same species. Comparative genomics research includes both developing computational tools and using the tools to analyze genomes for biological discoveries.

The practical applications of comparative genomics are many and its scientific impact profound [8]. For

* Corresponding author.
E-mail address: wei@nexusgenomics.com (L. Wei).

1532-0464/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S1532-0464(02)00506-3

Table 1
Comparison of the sizes of eight complete eukaryotae genome sequences and examples of complete bacteria and archaea genomes and eukaryotae draft genomes

Domain	Organism	Genome size (kbp)
Archaea	*Thermoplasma acidophilum*	1565
	Archaeoglobus fulgidus	2178
	Sulfolobus solfataricus	2992
	Methanosarcina acetivorans str. C2A	5751
Bacteria	*Salmonella typhi*	180
	Helicobacter pylori 26695	1668
	Haemophilus influenzae Rd	1830
	Escherichia coli K12	4639
Eucaryota	*Guillardia theta nucleomorph*	551
	Encephalitozoon cuniculi	2500
	Saccharomyces cerevisiae S288C	12,069
	Caenorhabditis elegans	97,000
	Arabidopsis thaliana	115,400
	Drosophila melanogaster	137,000
	Oryza sativa L. ssp. *indica* (draft)	420,000
	Homo sapiens (draft)	3,000,000

instance, animal models can be chosen as monitors for pathogenesis and therapy for human genetic diseases based on explicit gene orthology between the animal and human; effective treatment may result from analysis of the genetic differences between taxonomically related pathogen strains that have varying host responses. Scientific insights can be gained about species evolution in terms of gene birth and death, phylogeny of mammal orders, species origins, survival, and adaptations, just to name a few.

In the past five years there has been an explosion of computational and biological advances in the young field of comparative genomics. In this paper, we first briefly summarize the currently available computational tools for genome-scale sequence alignment. We then go into detail on how researchers have used comparative genomics to study the similarities and differences of organisms and strains. We focus on three important areas of comparative analysis: genome structure, coding regions, and noncoding regions. Finally, we discuss other related technologies and remaining technical challenges.

Before we start, it is important to define the differences between the terms homology and similarity. Homology implies common ancestry of two genes or gene products. Similarity is what we can measure from alignment of sequences or structures. Similarity may be used as evidence for homology, but does not necessarily imply homology. Most of this paper describes the analysis of similarity, which is a means towards studying homology.

2. Computational tools for genome-scale sequence alignment

The first step in comparative genomics analysis is often the alignment of two genome sequences. It is a technically challenging problem because of the large size of whole genomes (see Table 1), long insertions and deletions, large-scale rearrangement of genomic segments, and so on. Most traditional sequence alignment algorithms such as Smith-Waterman and BLASTN are no longer usable. In recent years a growing number of new algorithms have been developed for genome-scale alignment and visualization [9–17]. Table 2 lists some of the algorithms. Almost all of the alignment algorithms first identify large conserved sequence elements between the two genome sequences, and then generate the overall alignment. We refer readers interested in detailed discussions and comparisons of these alignment algorithms to several recent reviews [18–20].

Table 2
Examples of genome-scale alignment and visualization tools

Algorithms/tools	URL
BLASTN and MEGABLAST	http://www.ncbi.nlm.nih.gov/BLAST/
GLASS	http://crossspecies.lcs.mit.edu/
MUMmer	http://www.tigr.org/software/mummer/
PatternHunter	http://www.bioinformaticssolutions.com/products/ph.php
PipMaker	http://bio.cse.psu.edu/pipmaker/
VISTA	http://www-gsd.lbl.gov/vista/
WABA	http://www.cse.ucsc.edu/~kent/xenoAli/

3. Comparative analysis of genome structure

Analysis of the global structure of genomes, such as nucleotide composition, syntenic relationships, and gene ordering offer insight into the similarities and differences between genomes. Such comparisons provide information on the organization and evolution of the genomes, and highlight the unique features of individual genomes.

This analysis is now an indispensable part of any study on comparative sequence analysis. It has been applied to the analysis of the human genome [3,4], several yeast genomes [21], two nematode genomes [16], as well as the comparison of individual chromosomes from human and mouse [22–25].

The structure of different genomes can be compared at three levels: (1) overall nucleotide statistics, (2) genome structure at DNA level, and (3) genome structure at gene level.

3.1. Comparison of overall nucleotide statistics

Overall nucleotide statistics, such as genome size, overall (G+C) content, regions of different (G+C) content, and genome signature such as codon usage biases, amino acid usage biases, and the ratio of observed dinucleotide frequency and the expected frequency given random nucleotide distribution present a global view of the similarities and differences of the genomes.

For example, Mural et al. [25] noted that the mouse genome is about 10% smaller than the human genome in corresponding conserved regions, largely due to lower content of DNA repeats in mouse. Alm et al. [26] discovered that, although the two *Helicobacter pylori* strains J99 and 26695 have about the same overall (G+C) content, they each have several regions of different (G+C) content that are strain-specific; one of these regions is home to many strain-specific genes, which may indicate possible horizontal gene transfer. In another example, in their study of genome signature, Campbell et al. [27] found that plasmids and their hosts have highly similar genome signatures. Zeeberg [79] developed a Shannon Information Theoretic measure of synonymous codon usage and found a linear correlation between the information values of orthologous human and mouse sequences.

3.2. Comparison of genome structure at DNA level

Chromosomal breakage and exchange of chromosomal fragments are common mode of gene evolution. They can be studied by comparing genome structures at DNA level. Below we discuss four areas of comparative studies of particularly important biological significance.

3.2.1. Identification of conserved synteny and genome rearrangement events

The term "synteny" was originally coined to refer to gene loci on the same chromosome. It has since, however, taken on a new usage and often refers to two regions of two genomes that show considerable similarity of sequence and rough conservation of the order of genes in those regions, and thus are likely to be related by common descent. The terms "conserved synteny," "shared synteny," "syntenic," are sometimes used exchangeably with "synteny." Identification and analysis of syntenic regions provides information on the organization and evolution of genomes.

Synteny is detected either by identification of long, conserved sequence elements, or by comparison of conserved proteins using BLASTP, or both [9,25,28]. Important statistics about the syntenic regions include (1) the length of the regions and percentage of DNA sequence identity between conserved syntenic regions, (2) the percentage of genomic sequences that are within syntenic regions, (3) the distribution of these regions along the genomes, (4) the gene content, gene density, and gene order of conserved syntenic regions, and (5) content of DNA repeats. Description of interesting syntenic regions (especially those that are known to be associated with disease) is often provided. The result can be presented graphically showing the mapping of syntenic regions in corresponding genomes, from which genome rearrangement events can be identified such as fission, translocation, inversion, and transposition [8].

3.2.2. Analysis of breakpoints

Once syntenic regions are detected, one can obtain breakpoints (a.k.a. syntenic boundaries) between syntenic regions. Analysis of various genomic features of the breakpoints such as G+C content, gene density, and the density of various DNA repeats provides understanding of the evolution of genomes. For instance, Mural et al. [25] observed sharp discontinuity of features around some syntenic boundaries but not others. They hypothesized that syntenic boundaries that do not show sharp transitions in these various features may provide evidence for conservation of the ancestral pattern in the lineage.

3.2.3. Analysis of content and distribution of DNA repeats

DNA repeats (repetitive DNA sequences) are contained in most genomes. In human, 45% of the genome is made up of transposable elements, a type of DNA repeats. Analyzing the content and distribution of DNA repeats will shed light on their function. For example, Chureau et al. [29] analyzed the distribution of L1 elements (a type of repeats) in a region on the X chromosome in mouse, human and bovine, and found that in all three species there are more L1 elements in one strand of DNA than the other. Thus they hypothesized that L1 elements may have a potential function. A popular tool to analyze DNA repeats is RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker).

3.3 Comparison of genome structure at gene level

Chromosomal breakage and exchange of chromosomal fragments cause disruption of gene order. Therefore, gene order correlates with evolutionary

distance between genomes to a degree. Study of gene order has been done in various genome comparisons, including the comparison of two strains of *H. pylori* [26], yeast [30], two mycoplasm genomes [31], *Escherichia coli* vs. *H. influenzae* [32], and various prokaryotic genomes [33]. These studies analyze the conservation of gene order and conservation of relative orientation of gene pairs, and generate plots of positions of orthologs and paralogs in two species. These plots suggest hot spots of genome rearrangement. For instance, Seoighe et al. [30] compared local gene order in *Saccharomyces cerevisiae* and *Candida albicans* and calculated the percent of genes that are adjacent in both species and their order and orientation . By doing so, they showed that gene order is substantially different in these two yeasts, and that small reversals are prevalent in yeast gene order evolution.

The genome rearrangement problem is also a well-formulated problem to study the distance between genomes. Given genomes with distinct gene order, the problem of genome rearrangement is to find a series of rearrangements (e.g., reversal and transposition) to transform one genome into another. This has been shown to be a very hard computer science problem, but there are some computational tools available, such as GRIMM [34] (http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM/index.html).

4. Comparative analysis of coding regions

The comparative analysis of coding regions between different genomes typically involves the identification of gene-coding regions, comparison of gene content, and comparison of protein content. Recently there have also been a number of algorithms developed that use comparative genomics to aid function prediction of genes.

4.1. Identification of gene-coding regions

The analysis and comparison of the coding regions starts with, and is very dependent upon, the gene identification algorithm that is used to infer what portions of the genomic sequence actively code for genes. Gene identification is relatively straightforward in procaryote genomes, but remains a challenging problem in eucaryote genomes because of the high content of introns and intergenic regions, large repeat regions, alternative splicing, and so on. There are four basic approaches for gene identification, which are summarized in Table 3 [11,35–49]. A combination of multiple gene identification approaches are often used together in large-scale analysis to improve the overall accuracy [3,4].

4.2. Comparison of gene content

After the predicted gene set is generated, it is very interesting and important to compare the content of genes across genomes. The first statistics to compare is the estimated total number of genes in a genome, a statistics that has been made famous by the publication of the human genome [3,4]. Other statistics that elucidate the similarities and differences between the genomes include percentage of the genome that code for genes, distribution of coding regions across the genome (a.k.a. gene density), average gene length, codon usage, and so on.

Given the predicted gene set in different genomes, one can discover the percentage of genes that are common among the genomes, genes that are unique to each genome compared to the other genomes, and genes that are unique to each genome compared to known sequences in all other species in databases such as Genbank. This is often done using a pairwise sequence comparison tool such as BLASTN or TBLASTX [9]. For example, Mural et al. discovered from the com-

Table 3
Four basic categories of gene identification programs

Category	Algorithm	URL
Based on direct evidence of transcription	EST_GENOME	http://www.hgmp.mrc.ac.uk/Registered/Option/est_genome.html
	sim4	http://globin.cse.psu.edu/
Based on homology with known genes	PROCRUSTES	http://igs-server.cnrs-mrs.fr/igs/banbury/Procrustes-about.html
Statistical/ab initio approaches	Genscan	http://genes.mit.edu/GENSCAN.html
	Genie	http://www.fruitfly.org/seq_tools/genie.html
	FGENES	http://genomic.sanger.ac.uk/gf/Help/fgenes.html
	GeneMark	
	GeneMark.hmm	http://opal.biology.gatech.edu/GeneMark/
	HMMgene	http://www.cbs.dtu.dk/services/HMMgene/
	Glimmer	http://www.tigr.org/software/glimmer/glimmer.html
Using genome comparison	TwinScan	http://genes.cs.wustl.edu/
	Rosetta	http://crossspecies.lcs.mit.edu/
	SGP-1	http://soft.ice.mpg.de/sgp-1

parison of mouse chromosome 16 and syntenic regions in human genome that only 2% of all predicted mouse genes on chromosome 16 are specific to mouse, and similarly, only about 2.9% of human genes on the syntenic regions are specific to human [25].

4.3. Comparison of protein content (a.k.a. "comparative proteomics")

A second level of analysis that can be performed is to compare the set of gene products between the genomes, which has been termed "comparative proteomics." Assigning function to the protein sequences is a key first step. One usually starts with similarity-based seach tools such as BLAST against sequences with known functions in databases such as GenBank. Various schemes have been developed to derive the best functional assignment given the set of homologous genes found, beyond simply applying the annotation from the top match (subject to a scoring threshold). They often involve a voting scheme combining similarity level with textual analysis of the matching sequences' annotations [50]. Manual curation is still often used as the last step of function assignment for quality control.

It is important to compare the protein contents in critical pathways and important functional categories across genomes. The comparison allows one to identify specific pathways or functional categories that have high diversity across the genomes. Two widely used resources for pathways and functional categories are the KEGG pathway database and the Gene Ontology (GO) hierarchy [51,52]. Assigning proteins to KEGG or GO is a combination of automated assignment based on similarity to sequences with known KEGG or GO assignment and manual curation. It is important to compare and contrast the presence and abundance of proteins across genomes in various pathways and various components of pathways, and study factors such as whether one organism has significantly greater diversity in specific parts of a pathway than another organism. For example, Lin et al. [53] compared the restriction-modification systems in different strains of *H. pylori*.

It is common for genes to be replicated in a genome, and the replicated copies may take on similar or different function. By definition, orthologs refer to genes between different genomes that are evolved vertically from the same ancestral gene, whereas paralogs refer to genes within one genome derived from gene duplication. Sequence clustering algorithms have been applied to the set of protein sequences in a genome to find paralog families [26]. Comparison of corresponding paralog families across genomes provides evidence for finding the orthologous pairs. It allows one to ask important biological questions such as: Might organism A's adaptability to its environment come from a richer complement of related genes for some specific receptor [22]? Interesting statistics to compare include level of sequence identity between orthologous pairs across genome and between paralogous pairs within genome, number of replicated copies in corresponding paralog families, functions of the paralogs, and locations of members of paralog families across the genome.

4.4. Comparative genomics-based function prediction

About 40% of the predicted genes in newly sequenced genomes cannot be assigned any function based on sequence similarity to genes with known function. Recently, comparative genomic sequence analysis has been used to assist in the functional assignment of genes in a nonsimilarity-based manner. These comparative genomic approaches rely on the basic premise that genes that are functionally related are genes that are closely associated across genomes in some form. Here we discuss three of these methods, including (a) co-conservation across genomes, (b) conservation of gene clusters and genomic context across species, and (c) physical fusion of functionally linked genes across species. These nonsimilarity-based methods should be viewed as complementary, rather than as replacements to sequence-based methods for determining functional roles of genes. Note that while they do not rely directly on sequence similarity between a known and unknown gene for function assignment, they all critically employ sequence analysis to establish homologous and paralogous relationships of sequences across genomes.

4.4.1. Co-conservation across genomes

By observing the presence or absence of a gene in a genome across many genomes, one can establish a "phylogenetic profiles" for the gene. One might expect that functionally closely related genes, such as those involved in a metabolic pathway or structural complex, would tend to appear and disappear in genomes in a correlated manner. The profiles of these genes might also thus likely be highly correlated. Pellegrini et.al. used this premise to compare the "phylogenetic profiles" of 4290 proteins in *E. coli* against proteins in 16 other fully sequenced genomes [54]. They demonstrate that proteins with similar profiles are functionally linked, using the ribosomal protein RL7 as an example. More than half of the *E. coli* proteins with a phylogenetic profile similar to the ribosomal RL7 protein were found to have functions associated with the ribosome. Conversely, they demonstrate that groups of proteins known to be functionally linked (by keyword lookup) had many more pairs of phylogenetic "neighbors" on average than groups of randomly selected proteins.

While phylogenetic profiling is promising and often informative, not all functionally linked proteins have similar profiles leading to false negatives. The method also does not have a probabilistic accounting for the

strength of an association based on a "similar" phylogenetic profile to assist in distinguishing false positives. Related to this, the optimal parameters used to determine "similarity" are unclear.

4.4.2. Gene clusters

In prokaryotes, groups of functionally related genes tend to be located in close proximity to each other, and often in specific order, as exemplified by operons. Although gene order conservation beyond the level of operons is much less prevalent, conservation of clusters and gene order can be important indicators of function.

Several approaches have been used to determine functionally related "clusters" of genes. Overbeek et al. [55] use the constructs of a "pair of close bidirectional best hits" (PCBBH) and "pairs of close homologs" (PCHs) to represent pairs of genes that are closely conserved between two species and likely to be functionally related. Using PCBBHs and PCHs, 343 clusters of "role groups" were produced, and hundreds of hypothetical proteins were paired with proteins of known function. Wolf et al. [56] developed a program to construct gapped local alignments of conserved gene strings in two genomes . The alignment necessitates preservation of gene order, and mismatches (pairs of genes with no sequence similarity) were treated as gaps in scoring. The authors found that in most pairwise comparisons of genomes, $<10\%$ of genes in each genome belonged to conserved gene strings, although this ranged from $<5\%$ to 24% in closely related pairs of species. They conclude that gene order is poorly conserved among bacteria and archaea, but as a corollary, statistically conserved gene strings can be predicted to form operons. Using their methods, they were able to conservatively assign new functions or major clarifications for $\sim 4\%$ of 2422 analyzed Clusters of Orthologous Groups (COGs).

The constraint of gene proximity used in methods such as those described is not particularly strong, and can lead to many false positives. In addition, proteins that interact but are located far from each other will not be detected with these methods. Thresholds and cutoffs used in the methods are also empirically determined, and thus statistical validity of results is difficult to ascertain. Finally, the application of this methods to eukaryotes is limited as regulation of genes is much less tightly bound to genome structure.

4.4.3. Domain fusion analysis

It has been observed that certain pairs of interacting and functionally related proteins are fused in another organism into a single protein chain, referred to as a "Rosetta Stone" sequence or "composite" protein [57,58]. It is thus assumed that, if a composite protein is similar to two component proteins in another species, the two proteins are likely to be interacting and/or functionally related. Marcotte et al. [57] describe 6809 candidate protein-protein interactions in *E. coli* and 45,502 interactions in yeast based on this method. A cross validation of the results revealed 68% of pairs that shared at least one keyword in their SWISS-PROT annotations (vs. 15% random at baseline), confirmation of 6.4% of 724 pairs found in the Database of Interacting Proteins (experimentally derived protein interactions), and a 5% overlap with pairs identified using the phylogenetic profile method.

Certainly not all functionally related proteins have been fused into one protein in other organisms, and some fused proteins may have been lost in evolution. The extent of false positive results with this method is difficult to ascertain, but is likely to be higher for prediction of physical interaction of proteins than for functional association.

5. Comparative analysis of noncoding regions

Noncoding regions of the genome, which may comprise as much as 97% of the genome length such as in the human genome, gained a lot of attention in recent years because of its predicted role in regulation of transcription, DNA replication, and other biological functions [59,60]. However, identification of regulatory elements from the noncoding portion of a genome remains a challenge.

Comparative genomics has been used to greatly aid the identification of regulatory segments by comparing the genomic noncoding DNA sequences from diverse species to identify conserved regions (reviewed in [59,61]). This approach is based on the presumption that selective pressure causes regulatory elements to evolve at a slower rate than that of nonregulatory sequences in the noncoding regions. Indeed, human–rodent comparisons [62] indicate that only 19% of the human bases have greater than a 50% chance of being placed into an aligned block with a rodent base. Levy et al. [72] also demonstrated that conserved noncoding segments contain an enrichment of transcription factor binding sites when compared to the random sequence background.

This approach has been used successfully for the discovery of regulatory elements involved in the regulation of gene expression for many genes, including HBB (encoding β-globin) and BTK (encoding Bruton's tyrosine kinase) [63], IL 4,5,13 interleukins [64], stem cell leukemia gene (SCL) loci [65,66], cystic fibrosis transmembrane conductance regulator genes [67], and others [68–71]. It has been shown that the specificity for regulatory region detection increases significantly when more than two species are used in the comparative analysis [24,63,73,74]. It is based on the hypothesis that actively conserved human–mouse noncoding sequences, for example, will be present in a third mammal, whereas noncoding sequences that are similar between human

and mouse only because of an insufficient accumulation of random mutations will be absent in the third mammal. By comparing human and mouse sequences, Frazer et al. found that one-half of the human–mouse conserved noncoding sequence was also conserved in a third mammal, the dog [24].

One more direction towards high-throughput identification of regulatory sites is to combine searching transcription factor binding sites database with identifying highly conserved sequence regions, using global sequence alignment of syntenic regions and clustering [75]. This fast procedure reduces the number of predicted transcription factor binding sites by several orders of magnitude and thus increases the specificity significantly.

6. Discussion

There are many other exciting technologies in comparative genomics than what we have covered so far. To give two examples, oligonucleotide array technology was used in the identification of conserved noncoding elements from human chromosome 21 [24], and chromosome painting has been used to demonstrate gross genomic rearrangements [76]. There are also other important biological problems for which comparative genomics has played key roles, such as noncoding RNA gene detection [77,78]. Many challenges, however, still remain. For example, there is a lack of efficient multiple-sequence alignment algorithms for genome-scale sequences. There is a strong need for rigorous modeling and evaluation of the statistical significance of regulatory region predictions. Manual curation is often still required as the last step of recognition of genome rearrangement events, assignment of gene function, and prediction of regulatory region. Though expert opinion will always be of significant value, computational methods are needed to automate these steps as much as possible.

Comparative genomics is undoubtedly one of the most promising scientific fields today, and we anticipate more and more exciting technical advances and biological discoveries in the future.

Acknowledgments

Yueyi Liu thanks Diane Oliver, Jeff Chang, and Soumya Raychaudhuri for their helpful comments.

References

[1] Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. Science 1995;269:496–512.

[2] NCBI. Genome Resources. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome.

[3] Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304–51.

[4] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

[5] Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 2002;296:79–92.

[6] Goff SA, Ricke D, Lan TH, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 2002;296:92–100.

[7] Clark MS. Comparative genomics: the key to understanding the Human Genome Project. Bioessays 1999;21:121–30.

[8] O'Brien SJ, Menotti-Raymond M, Murphy WJ, et al. The promise of comparative genomics in mammals. Science 1999;286:458–62, 479–481.

[9] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.

[10] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[11] Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res 2000;10:950–8.

[12] Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 2002;30:2478–83.

[13] Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res 1999;27:2369–76.

[14] Schwartz S, Zhang Z, Frazer KA, et al. PipMaker–a web server for aligning two genomic DNA sequences. Genome Res 2000;10:577–86.

[15] Mayor C, Brudno M, Schwartz JR, et al. VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 2000;16:1046–7.

[16] Kent WJ, Zahler AM. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. Genome Res 2000;10:1115–25.

[17] Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics 2002;18:440–5.

[18] Dubchak I, Pachter L. The computational challenges of applying comparative-based computational methods to whole genomes. Brief Bioinform 2002;3:18–22.

[19] Miller W. Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics 2001;17:391–7.

[20] Wiehe T, Guigo R, Miller W. Genome sequence comparisons: hurdles in the fast lane to functional genomics. Brief Bioinform 2000;1:381–8.

[21] Cliften PF, Hillier LW, Fulton L, et al. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Res 2001;11:1175–86.

[22] Dehal P, Predki P, Olsen AS, et al. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. Science 2001;293:104–11.

[23] Deloukas P, Matthews LH, Ashurst J, et al. The DNA sequence and comparative analysis of human chromosome 20. Nature 2001;414:865–71.

[24] Frazer KA, Sheehan JB, Stokowski RP, et al. Evolutionarily conserved sequences on human chromosome 21. Genome Res 2001;11:1651–9.

[25] Mural RJ, Adams MD, Myers EW, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 2002;296:1661–71.

[26] Alm RA, Ling LS, Moir DT, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 1999;397:176–80.

[27] Campbell A, Mrazek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci USA 1999;96:9184–9.
[28] Salse J, Piegu B, Cooke R, Delseny M. Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res 2002;30:2316–28.
[29] Chureau C, Prissette M, Bourdet A, et al. Comparative sequence analysis of the x-inactivation center region in mouse, human, and bovine. Genome Res 2002;12:894–908.
[30] Seoighe C, Federspiel N, Jones T, et al. Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci USA 2000;97:14433–7.
[31] Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res 1997;25:701–12.
[32] Tatusov RL, Mushegian AR, Bork P, et al. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr Biol 1996;6: 279–91.
[33] Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 2001;17:10–3.
[34] Tesler G. GRIMM: genome rearrangements web server. Bioinformatics 2002;18:492–3.
[35] Mott R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 1997;13:477–8.
[36] Bailey Jr LC, Searls DB, Overton GC. Analysis of EST-driven gene annotation in human genomic sequence. Genome Res 1998;8:362–76.
[37] Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 1998;8:967–74.
[38] Birney E, Thompson JD, Gibson TJ. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res 1996;24:2730–9.
[39] Novichkov PS, Gelfand MS, Mironov AA. Gene recognition in eukaryotic DNA by comparison of genomic sequences. Bioinformatics 2001;17:1011–8.
[40] Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced sequence alignment. Proc Natl Acad Sci USA 1996;93:9061–6.
[41] Krogh A. Two methods for improving performance of an HMM and their application for gene finding. Proc Int Conf Intell Syst Mol Biol 1997;5:179–86.
[42] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94.
[43] Reese MG, Kulp D, Tammana H, Haussler D. Genie–gene finding in *Drosophila melanogaster*. Genome Res 2000;10:529–38.
[44] Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 1998;26:1107–15.
[45] Isono K, McIninch JD, Borodovsky M. Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program GeneMark. DNA Res 1994;1:263–9.
[46] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27:4636–41.
[47] Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Res 1998;26:544–8.
[48] Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinformatics 2001; 17(Suppl 1):S140–8.
[49] Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R. SGP-1: prediction and validation of homologous genes based on sequence alignments. Genome Res 2001;11:1574–83.
[50] Zhou Y, Huang GM, Wei L. UniBlast: a system to filter, cluster, and display BLAST results and assign unique gene annotation. Bioinformatics 2002;18(9):1268–9.
[51] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[52] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.
[53] Lin LF, Posfai J, Roberts RJ, Kong H. Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. Proc Natl Acad Sci USA 2001;98:2740–5.
[54] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–8.
[55] Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999;96:2896–901.
[56] Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 2001;11:356–72.
[57] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999;285:751–3.
[58] Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
[59] Hardison RC. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 2000;16:369–72.
[60] Meisler MH. Evolutionarily conserved noncoding DNA in the human genome: how much and what for? Genome Res 2001;11:1617–8.
[61] Pennacchio LA, Rubin EM. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2001;2: 100–9.
[62] Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Human-mouse genome comparisons to locate regulatory sites. Nat Genet 2000;26:225–8.
[63] Hardison RC, Oeltjen J, Miller W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res 1997;7:959–66.
[64] Loots GG, Locksley RM, Blankespoor CM, et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 2000;288:136–40.
[65] Gottgens B, Barton LM, Gilbert JG, et al. Analysis of vertebrate SCL loci identifies conserved enhancers. Nat Biotechnol 2000;18:181–6.
[66] Gottgens B, Gilbert JG, Barton LM, et al. Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. Genome Res 2001;11: 87–97.
[67] Ellsworth RE, Jamison DC, Touchman JW, et al. Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. Proc Natl Acad Sci USA 2000;97:1172–7.
[68] Lund J, Chen F, Hua A, et al. Comparative sequence analysis of 634 kb of the mouse chromosome 16 region of conserved synteny with the human velocardiofacial syndrome region on chromosome 22q11.2. Genomics 2000;63:374–83.
[69] Leung JY, McKenzie FE, Uglialoro AM, et al. Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters. Proc Natl Acad Sci USA 2000;97:6614–8.

[70] Jareborg N, Birney E, Durbin R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res 1999;9:815–24.
[71] Wu Q, Zhang T, Cheng JF, et al. Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. Genome Res 2001;11:389–404.
[72] Levy S, Hannenhalli S, Workman C. Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics 2001;17:871–7.
[73] Dubchak I, Brudno M, Loots GG, et al. Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res 2000;10:1304–6.
[74] Mouchel N, Tebbutt SJ, Broackes-Carter FC, et al. The sheep genome contributes to localization of control elements in a human gene with complex regulatory mechanisms. Genomics 2001;76:9–13.
[75] Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. For comparative sequence-based discovery of functional transcription factor binding sites. Genome Res 2002;12:832–9.
[76] Scherthan H, Cremer T, Arnason U, Weier HU, Lima-de-Faria A, Fronicke L. Comparative chromosome painting discloses homologous segments in distantly related mammals. Nat Genet 1994;6:342–7.
[77] Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinform 2001;2:8.
[78] Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol 2001;11:1369–73.
[79] Zeeberg B. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. Genome Res 2002;12(6):944–55.

Statement of ownership, management, and circulation required by the Act of October 23, 1962, Section 4369, Title 39, United States Code: of

JOURNAL OF BIOMEDICAL INFORMATICS

Published bimonthly by Academic Press, 6277 Sea Harbor Drive, Orlando, FL 32887-4900. Number of issues published annually: 6. Editor: Edward Shortliffe, Dept Med Informatics, Columbia University NY Presbyterian, Vanderbilt Clinic Bldg., 5th Floor, New York NY 10032-3784.
Owned by Academic Press, 525 B Street, Suite 1900, San Diego, CA 92101-4495. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, and other securities: None.
Paragraphs 2 and 3 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages, or other securities of the publishing corporation have been included in paragraphs 2 and 3 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.
Total no. copies printed: average no. copies each issue during preceding 12 months: 1300; single issue nearest to filing date: 925. Paid circulation (a) to term subscribers by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 686; single issue nearest to filing date: 338. (b) Sales through agents, news dealers, or otherwise: average no. copies each issue during preceding 12 months: 259; single issue nearest to filing date: 317. Free distribution (a) by mail: average no. copies each issue during preceding 12 months: 39; single issue nearest to filing date: 39. (b) Outside the mail: average no. copies each issue during preceding 12 months: 8; single issue nearest to filing date: 8. Total no. of copies distributed: average no. copies each issue during preceding 12 months: 992; single issue nearest to filing date: 702. Percent paid and/or requested circulation: average percent each issue during preceding 12 months: 95%; single issue nearest to filing date: 93%.

(Signed) Stephanie Smith, Assoc. Manager, Global Sales Operations