Web-based resources for comparative genomics

Xun Gu[1]* *and Zhixi Su*[1,2]

[1] Department of Genetics, Development and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA
[2] James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, China
**Correspondence to*: Tel: +1 515 294 8075; Fax: +1 515 294 8457; E-mail: xgu@iastate.edu

Date received (in revised form): 19th May 2005

Abstract

The available web-based genome data and related resources provide great opportunities for biomedical scientists to identify functional elements in a particular genome region or to explore the evolutionary pattern of genome dynamics. Comparative genomics is an indispensable tool for achieving these goals. Because of the broad scope of comparative genomics, it is difficult to address all of its aspects in this short survey. A few currently 'hot' topics have therefore been selected and a brief review of the availability of web-based databases and software is given.

Keywords: *comparative genomics, software, web-based database*

Genome databases for comparative genomics

Usually, genome-wide databases (see Table 1) change rapidly, both in their internal implementation and in the datasets recorded. This paper briefly reviews two severs recently made public, which researchers should find valuable for obtaining a wealth of useful information. The genome alignment and annotation database (GALA)[1] provides access to information on genes (known and predicted), gene ontology, expression patterns, genome alignments and conserved transcription factor binding sites predicted by the TRANSFAC weight matrix that can be estimated from the known binding sites to show the sequence signature.[2] For example, given a set of genes expressed in a particular tissue, GALA is able to identify all of the predicted binding sites for one or more transcription factors of interest that are all conserved in mammals. EnsMart is a branch of the Ensembl project,[3] which integrates data from Ensembl and several other resources, using a 'warehouse star-schema' with central biological objects (eg genes or single nucleotide polymorphisms) connected to a set of satellite tables, such as disease, transcript and protein family (PFAM) attributes. Thus, EnsMart provides users with fast and effective access to deep data in and around genes.

Multi-genome alignment and gene prediction

Genome-wide alignment servers for two closely related species are available on the web. The BLAST,[4,5] implemented at the National Center for Biotechnology Information (NCBI), is the most frequently used suite of tools. Several servers were specially designed to align two or more long genomic sequences at high sensitivity while detecting common rearrangements or duplications — for example, PipMaker,[6] MultiPipMaker,[7] zPicture,[8] VISTA[9] and MAVID.[10] These servers are suitable for species such as those from different mammalian orders. Several pipelines have been designed for mammalian genome alignment.[11–13] For more distant species, or ancient paralogous genes, different alignment methods should be recommended. One major application is to look for common motifs in the upstream regions of co-expressed genes. Two examples of these approaches are multiple expectation maximisation for motif elicitation (MEME) and Gibbs sampling.[14–16]

One application of multi-genome alignment is to improve the efficiency of gene finding. ROSETTA reconstructs co-linear gene structures from global alignments and defines exons as sub-sequences bounded by splice sites.[17] Syntenic Gene Prediction version 1 (SGP18) reconstructs genes from a collection of local alignments between two sequences,[18] while

Table 1. Websites for tools and databases useful for comparative genomics.

Tool or database	Website
NCBI	http://www.ncbi.nlm.nih.gov
ENSEMBL	http://www.ensembl.org
UCSC Genome Browser	http://genome.uscs.edu/
EnsMart	http://www.ensembl.org/Multi/martview
NCBI BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
WU-BLAST	http://blast.wustl.edu/
GALA	http://gala.cse.psu.edu/
PipMaker and MultiPipMaker	http://bio.cse.psu.edu/pipmaker/
zPicture	http://zpicture.dcode.org/
VISTA	http://www-gsd.lbl.gov/vista/
MAVID	http://baboon.math.berkeley.edu/mavid/
MEME	http://meme.sdsc.edu
GLASS and ROSETTA	http://crossspecies.lcs.mit.edu/
SGP2	http://genome.imim.es/software/sgp2/
TWINSCAN	http://genes.cs.wustl.edu/query.html
GeneID	http://www1.imim.es/geneid.html
DOUBLESCAN	http://www.sanger.ac.uk/Software/analysis/doublescan/
TRED	http://rulai.cshl.edu/TRED
RNAdb	http://research.imb.uq.edu.au/rnadb/
NONCODE	http://noncode.bioinfo.org.cn
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html
DIVERGE	http://xgu.zool.iastate.edu
Mgenome	http://xgu.zool.iastate.edu
GRIMM	http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM/
GRAPPA	http://www.cs.unm.edu/~moret/GRAPPA/
TRANSFAC	http://www.gene-regulation.de/
FootPrinter and PhyME	http://bio.cs.washington.edu/software.html
MSARI	http://theory.csail.mit.edu/MSARi/
RNAZ	http://www.tbi.univie.ac.at/~wash/RNAz/

SGP2 assesses the reliability of gene models predicted by GeneID,[19] a conventional gene predictor.[20] Similarly, TWINSCAN represents a direct extension of the Genscan algorithm that integrates conservation information between two sequences.[21–23] DOUBLESCAN uses a pair hidden Markov model (Pair HMM) to reconstruct gene structures from a series of local alignments created with BLAST.[4,24]

Evolutionary approaches to protein function detection

Phylogenetic analysis by maximum likelihood (PAML) is a software package that includes a wealth of methods for statistically testing the evolutionary pattern of coding sequences, which can be used for one functional detection and prediction

of proteins.[25] For instance, PAML is able to estimate ω, the ratio of the non-synonymous rate to the synonymous rate at each amino acid residue along the lineages of a given phylogenetic tree. DIVERGE is a program for studying one functional divergence of a protein family by detecting site-specific changes in the evolutionary rate using a multiple alignment of amino acid sequences for a given phylogenetic tree.[26,27] It first conducts a statistical test for site-specific rate shifts along the tree and predicts candidate amino acid residues responsible for functional divergence based on posterior analysis. These results can then be mapped on the three-dimensional protein structure, if available.

Multiple genome rearrangement by signed reversal

For comparative gene mapping, it is important to reconstruct the ancestral gene orders for given current genomes. Mathematically, it becomes a problem of signed reversals — that is, how the genomes evolve from a common ancestral genome based on signed reversal of genes or gene sets. Since this problem is now-deterministic polynomial-time hard (NP-hard),[28] most work is focused on heuristic algorithms for reconstructing the gene order of ancestral genomes. Sankoff *et al.*[29] searched for the optimal ancestral genome for a median problem upon a grid. Bourque and Pevzner[30] designed the model generative reasoning (MGR) algorithm to reconstruct ancestral genomes using a greedy-split strategy. Wu and Gu[31,32] improved the searching accuracy by using a nearest path search algorithm; they developed a neighbour-perturbing algorithm to reconstruct the optimal gene order of ancestral genomes.

Comparative microarray analysis

Because of the limited data available, there are only a few case studies for interspecies microarray analysis. One good example is for the human–chimpanzee expression profile comparisons in the brain and liver.[33,34] Gu[35] developed a statistical framework for studying expression divergence between duplicate genes, which can also be used to infer the ancestral expression profiles when the phylogeny of duplicate genes is known. To facilitate application of these models to expression and genomic data, Gu *et al.*[36] defined an additive expression distance between duplicate genes, measured by the average of squared expression differences. They analysed yeast gene families using a multi-microarray dataset and found a more than ten-fold increase in the rate of expression evolution immediately following gene duplication.

Identification of functional non-coding elements by comparative genomics

Although the majority of eukaryote genomes are non-coding regions and were previously regarded as 'junk DNA', recent studies have indicated that non-coding regions harbour important functional elements such as *cis*-regulatory modules.[37,38] Computational detection of these functional non-coding elements has been extremely challenging. It has been recognised that comparative genomics may be a promising approach to solving this problem. 'Phylogenetic footprinting' focuses on the discovery of novel regulatory elements based on the sequence conservation among a set of orthologous non-coding regions.[39] Using this method, many successful motif discovery programs have been developed; for example, Gibbs sampler,[40] MEME,[41] Consensus,[42] AlignAce,[43] ANN-Spec,[44] FootPrinter[45] and PhyMe.[46] For non-coding RNA elements, many tools have been developed to identify the evolutionary conservation of secondary structures, such as QRNA,[47] DDBRNA,[48] MSARI,[49] and RNAZ.[50] The development of these tools serves as compelling evidence for biologically relevant non-coding RNAs function. In addition, some databases of functional non-coding elements are also available; for example, TRED,[51] RNAdb[52] and NONCODE.[53]

Conclusion

In summary, this paper has briefly reviewed the web-based resources for comparative genomics. Given that substantial resources are available, the challenge in fact turns on how to transfer the explosion in genomic data to biological knowledge. The internet has substantially facilitated the transition process but progress depends on the development of new ideas and analysis pipelines that combine many approaches, including comparative genomics.

Acknowledgments

This work was supported by an NIH grant to X.G. and the NSFC Overseas Outstanding Young Investigator Award (China) to X.G.

References

1. Giardine, B.M., Elnitski, L., Riemer, C. *et al.* (2003), 'GALA, a database for genomic sequence alignments and annotations', *Genome Res.* Vol. 13, pp. 732–741.
2. Wingender, E., Chen, X., Fricke, E. *et al.* (2001), 'The TRANSFAC system on gene expression regulation', *Nucleic Acids Res.* Vol. 29, pp. 281–283.
3. Kasprzyk, A., Keefe, D., Smedley, D. *et al.* (2003), 'EnsMart—A generic system for fast and flexible access to biological data', *Genome Res.* Vol. 14, pp. 160–169.
4. Altschul, S.F., Gish, W., Miller, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.* Vol. 215, pp. 403–410.
5. Altschul, S.F., Madden, T.L., Schaffer, A.A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.* Vol. 25, pp. 3389–3402.
6. Schwartz, S., Zhang, Z., Frazer, K.A. *et al.* (2000), 'PipMaker—A web server for aligning two genomic DNA sequences', *Genome Res.* Vol. 10, pp. 577–586.

7. Schwartz, S., Elnitski, L., Li, M. *et al.* (2003), 'MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences', *Nucleic Acids Res.* Vol. 31, pp. 3518–3524.
8. Ovcharenko, I., Loots, G.G., Hardison, R.C. *et al.* (2004), 'zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles', *Genome Res.* Vol. 14, pp. 472–477.
9. Mayor, C., Brudno, M., Schwartz, J.R. *et al.* (2000), 'VISTA: Visualizing global DNA sequence alignments of arbitrary length', *Bioinformatics* Vol. 16, pp. 1046–1047.
10. Bray, N. and Pachter, L. (2003), 'MAVID multiple alignment server', *Nucleic Acids Res.* Vol. 31, pp. 3525–3526.
11. Brudno, M., Do, C.B., Cooper, G.M. *et al.* (2003), 'LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA', *Genome Res.* Vol. 13, pp. 721–731.
12. Couronne, O., Poliakov, A., Bray, N. *et al.* (2003), 'Strategies and tools for whole-genome alignments', *Genome Res.* Vol. 13, pp. 73–80.
13. Schwartz, S., Kent, W.J., Smit, A. *et al.* (2003), 'Human-mouse alignments with Blastz', *Genome Res.* Vol. 13, pp. 103–105.
14. Bailey, T.L. and Elkan, C. (1995), 'The value of prior knowledge in discovering motifs with MEME', *Proc. Int. Conf. Intell. Syst. Mol. Biol.* Vol. 3, pp. 21–29.
15. Schug, J. and Overton, G.C. (1997), 'Modeling transcription factor binding sites with Gibbs sampling and minimum description length encoding', *Proc. Int. Conf. Intell. Syst. Mol. Biol.* Vol. 5, pp. 268–271.
16. Thompson, W., Rouchka, E.C. and Lawrence, C.E. (2003), 'Gibbs Recursive Sampler: Finding transcription factor binding sites', *Nucleic Acids Res.* Vol. 31, pp. 3580–3585.
17. Batzoglou, S., Pachter, L., Mesirov, J.P. *et al.* (2000), 'Human and mouse gene structure: Comparative analysis and application to exon prediction', *Genome Res.* Vol. 10, pp. 950–958.
18. Wheeler, D.L., Church, D.M. and Lash, A.E. (2002), 'Database resources of the National Center for Biotechnology Information: Update', *Nucleic Acids Res.* Vol. 30, pp. 13–16.
19. Parra, G., Agarwal, P., Abril, J.F. *et al.* (2003), 'Comparative gene prediction in human and mouse', *Genome Res.* Vol. 13, pp. 108–117.
20. Guigo, R. (1998), 'Assembling genes from predicted exons in linear time with dynamic programming', *J. Comput. Biol.* Vol. 5, pp. 681–702.
21. Korf, I., Flicek, P., Duan, D. and Brent, M.R. (2001), 'Integrating genomic homology into gene structure prediction', *Bioinformatics* Vol. 17(Suppl. 1), pp. S140–S148.
22. Burge, C. (1997), 'Identification of genes in human genomic DNA', PhD thesis, Stanford University, Stanford, CA.
23. Burge, C. and Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *J. Mol. Biol.* Vol. 268, pp. 78–94.
24. Meyer, I.M. and Durbin, R. (2002), 'Comparative *ab initio* prediction of gene structures using pair HMMs', *Bioinformatics* Vol. 18, pp. 1309–1318.
25. Yang, Z. (1997), 'PAML: A program package for phylogenetic analysis by maximum likelihood', *Comput. Appl. Biosci.* Vol. 13, pp. 555–556.
26. Gu, X. and Vander Velden, K. (2002), 'DIVERGE: Phylogeny-based analysis for functional-structural divergence of a protein family', *Bioinformatics* Vol. 18, pp. 500–501.
27. Gu, X. (1999), 'Statistical methods for testing functional divergence after gene duplication', *Mol. Biol. Evol.* Vol. 16, pp. 1664–1674.
28. Caprara, A. (1999), 'Formulations and hardness of multiple sorting by reversals', Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB'99), ACM Press, New York, NY.
29. Sankoff, D., Sudaram, G. and Kececioglu, J. (1996), 'Steiner points in the space of genome rearrangements', *Int. J. Found. Comput. Sci.* Vol. 7, pp. 1–9.
30. Bourque, G. and Pevzner, P.A. (2002), 'Genome-scale evolution: reconstructing gene orders in the ancestral species', *Genome Res.* Vol. 12, pp. 26–36.
31. Wu, S. and Gu, X. (2002), 'Multiple genome rearrangement by reversals', *Pac. Symp. Biocomput.* Vol. 7, pp. 259–270.
32. Wu, S. and Gu, X. (2003), 'Algorithms for multiple genome rearrangement by signed reversals', *Pac. Symp. Biocomput.* Vol. 8, pp. 363–374.
33. Enard, W., Khaitovich, P., Klose, J. *et al.* (2002), 'Intra- and interspecific variation in primate gene expression patterns', *Science* Vol. 296, pp. 340–343.
34. Gu, J. and Gu, X. (2003), 'Induced gene expression in human brain after the split from chimpanzee', *Trends Genet.* Vol. 19, pp. 63–65.
35. Gu, X. (2004), 'Statistical framework for phylogenomic analysis of gene family expression profiles', *Genetics* Vol. 167, pp. 531–542.
36. Gu, X., Zhang, Z. and Huang, W. (2005), 'Rapid evolution of expression and regulatory divergences after yeast gene duplication', *Proc. Natl. Acad. Sci. USA* Vol. 102, pp. 707–712.
37. Dermitzakis, E.T., Reymond, A., Lyle, R. *et al.* (2002), 'Numerous potentially functional but non-genic conserved sequences on human chromosome 21', *Nature* Vol. 420, pp. 578–582.
38. Gibbs, W.W. (2003), 'The unseen genome: Gems among the junk', *Sci. Am.* Vol. 289, pp. 26–33.
39. Tagle, D.A., Koop, B.F., Goodman, M. *et al.* (1988), 'Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints', *J. Mol. Biol.* Vol. 203, pp. 439–455.
40. Lawrence, C.E., Altschul, S.F., Boguski, M.S. *et al.* (1993), 'Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment', *Science* Vol. 262, pp. 208–214.
41. Bailey, T.L. and Elkan, C. (1995), 'Unsupervised learning of multiple motifs in biopolymers using expectation maximization', *Mach. Learn.* Vol. 21, pp. 51–80.
42. Hertz, G.Z., Stormo, G.D. (1999), 'Identifying DNA and protein patterns with statistically significant alignments of multiple sequences', *Bioinformatics* Vol. 15, pp. 563–577.
43. Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998), 'Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation', *Nat. Biotechnol.* Vol. 16, pp. 939–945.
44. Workman, C.T. and Stormo, G.D. (2000), 'ANN-Spec: A method for discovering transcription factor binding sites with improved specificity', *Pac. Symp. Biocomput.* Vol. 5, pp. 467–478.
45. Blanchette, M. and Tompa, M. (2003), 'FootPrinter: A program designed for phylogenetic footprinting', *Nucleic Acids Res.* Vol. 31, pp. 3840–3842.
46. Sinha, S., Blanchette, M. and Tompa, M. (2004), 'PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences', *BMC Bioinformatics* Vol. 5, p. 170.
47. Rivas, E. and Eddy, S.R. (2001), 'Noncoding RNA gene detection using comparative sequence analysis', *BMC Bioinformatics* Vol. 2, p. 8.
48. di Bernardo, D., Down, T. and Hubbard, T. (2003), 'ddbRNA: Detection of conserved secondary structures in multiple alignments', *Bioinformatics* Vol. 19, pp. 1606–1611.
49. Coventry, A., Kleitman, D.J. and Berger, B. (2004), 'MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure', *Proc. Natl. Acad. Sci. USA* Vol. 101, pp. 12102–12107.
50. Washietl, S., Hofacker, I.L. and Stadler, P.F. (2005), 'Fast and reliable prediction of noncoding RNAs', *Proc. Natl. Acad. Sci. USA* Vol. 102, pp. 2454–2459.
51. Zhao, F., Xuan, Z., Liu, L. and Zhang, M.Q. (2005), 'TRED: A Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies', *Nucleic Acids Res.* Vol. 33, pp. D103–D107.
52. Pang, K.C., Stephen, S., Engstrom, P.G. *et al.* (2005), 'RNAdb — A comprehensive mammalian noncoding RNA database', *Nucleic Acids Res.* Vol. 33, pp. D125–D130.
53. Liu, C., Bai, B., Skogerbo, G. *et al.* (2005), 'NONCODE: An integrated knowledge database of non-coding RNAs', *Nucleic Acids Res.* Vol. 33, pp. D112–D115.