BIOINFORMATICS REVIEW

Vol. 17 no. 5 2001
Pages 391–397

Comparison of genomic DNA sequences: solved and unsolved problems

Webb Miller

Department of Computer Science and Engineering, Penn State, University Park, PA 16802, USA

Received on November 31, 2000; revised and accepted on January 5, 2001

ABSTRACT

Motivation: The DNA sequences of entire genomes are being determined at a rapid rate. Whereas initial genome sequencing efforts were for organisms chosen to be widely spaced in the tree of life, there is a growing emphasis on projects to sequence a species that is sufficiently similar to an already-sequenced species to allow direct comparison of those two DNA sequences. This and other changes in genome sequencing strategies have created a strong need for new methods to compare genomic sequences.

Results: We sketch the current state of software for comparing genomic DNA sequences and outline research directions that we believe are likely to result in important advances in practice.

Contact: webb@cse.psu.edu

INTRODUCTION

The art and science of comparing genomic DNA sequences have changed dramatically in the last two or three years, with respect to the anticipated rate, nature, and utilization of the data. It is incumbent on the bioinformatics community to understand these changes and to respond appropriately. This paper sketches one worker's perception of the field's current status and predicts several fruitful research directions. The reader should keep in mind that the opinions expressed here are highly personal and that no attempt has been made to give an exhaustive survey of the field.

Perhaps the most striking change in this area is in the sheer volume of the available data. Original plans called for completion of the human genome sequence by the year 2005. Along the way, genomic sequences were to be determined for one bacterium, a yeast, a worm, and a fly, but little mention was made of possible sequencing projects for the genomes of other vertebrates. Surprisingly, it now seems certain that by 2005 we will have genomic sequences for human, mouse, rat, and a couple of fishes, with additional vertebrate genomes also under consideration. Even two years ago, such a bounty of data in this time frame seemed out of the question.

The anticipated nature of the genomic sequence data has changed, too. Early discussions centered around the goal of producing a complete sequence of extremely high accuracy. In practice, however, the 'finished' versions of human chromosomes 21 and 22 contain gaps where the data could not be acquired, and the biologists eagerly awaiting data for the other chromosomes are currently working with 'draft' sequence data, which frequently consists of pieces whose relative order and orientation are difficult to determine. Dealing with such incomplete data naturally places new demands upon software tools, particularly when two of these sequences are being compared, though early studies suggest that many of the difficulties can be overcome (e.g. Onyango *et al.*, 2000). Indeed, one reasonable strategy is to finish only one genomic sequence and to merely sample the genomes of closely related species (e.g. McClelland *et al.*, 2000), which can be done at a fraction of the cost of finished sequences.

A third shift is that the types of analyses one wants to perform on these data have come into clearer focus, and in some cases are strikingly different than what was anticipated just a few years ago. Initial discussions of the value of human–mouse alignments generally focused on their effectiveness for identifying non-coding regions with an important biological function, particularly those involved in regulating gene transcription (e.g. Hardison and Miller, 1993; Koop and Hood, 1994; Duret and Bucher, 1997; Hardison *et al.*, 1997). Once substantial human–mouse datasets began to accumulate it became clear that human–mouse genomic sequence comparisons will be very valuable for finding protein coding regions (Makalowski *et al.*, 1996; Ansari-Lari *et al.*, 1998; Jang *et al.*, 1999). Also, realistic analyses (e.g. Dunham *et al.*, 1999; Guigó *et al.*, 2000) of the effectiveness of alternative gene prediction methods underscored the need for improved prediction accuracy. These converging observations have substantially accelerated the mouse sequencing efforts and created the need for novel sequence-comparison tools.

Thus, a contemporary view might be that interspecies

genomic comparisons will first be used to aid identification of all protein-coding regions. Subsequently, and probably extending over a much longer time period, these comparisons will be used to locate signals that regulate gene transcription, to understand the mechanisms and tempo of genome evolution, and to identify hitherto unimagined segments that modulate the structure and function of the genome.

Here we give a personal opinion of the state and desirable direction of software for comparing genomic DNA sequences. A disproportionate fraction of the discussion concerns pairwise alignment algorithms; this reflects the relatively thorough exploration of those methods to date, rather than an urgency for new developments. Indeed, the need for work in certain other areas is more pressing, precisely because little work on them has been completed and little is known about how research might best proceed.

In summary, this paper identifies the following immediate needs (in no particular order).

(1) Improved software that aligns two genomic sequences and has a rigorous statistical basis.

(2) An industrial-strength gene prediction system that effectively combines genomic sequence comparisons, intrinsic sequence properties, and results from searching databases of proteins sequences and ESTs.

(3) Reliable and automatic software for aligning three or more genomic sequences.

(4) Better methods for displaying and browsing genomic sequence alignments.

(5) Improved datasets and protocols for evaluating the correctness and performance of genomic alignment software.

Of course, the hope is that the fruits of these efforts will quickly be placed in the hands of biologists, in the form of network servers and/or portable software.

PAIRWISE ALIGNMENT ALGORITHMS

Alignment of two genomic sequences poses problems not well addressed by earlier alignment programs, which were typically designed for protein sequences. Most such programs are incapable of producing accurate long alignments, and may have other deficiencies for genomic sequences. For instance, the Blastn program does not permit alignment scores that distinguish transitions from transversions, much less ones that model, e.g. nucleotide substitution patterns that depend on the isochore.

A number of newer tools are aimed at comparing two genomic DNA sequences. Examples include MUMmer (Delcher *et al.*, 1999), DBA (Jareborg *et al.*, 1999), GLASS (Batzoglou *et al.*, 2000), WABA (Kent and Zahler, 2000) and Dialign (Morgenstern *et al.*, 1998; Göttgens *et al.*, 2001). These programs use a variety of different methods; a detailed comparison of their performances would be quite useful, but is beyond the scope of this paper.

We are most familiar with the strengths and weaknesses of the alignment program used by the PipMaker network server (Schwartz *et al.*, 2000), and will limit our detailed comments to it. That program, called 'blastz', uses an approach similar to the gapped blast program (Altschul *et al.*, 1997). Instead of the widely used notion of locally optimal alignment proposed by Smith and Waterman (1981), blastz uses the 'X-drop' approach (Zhang *et al.*, 1998), which we prefer for reasons given by Zhang *et al.* (1999).

We emphasize that blastz calculates *local* alignments; i.e. given two sequences, it produces a set of alignments that individually cover only a portion of each sequence. We believe that this is necessary for a general purpose genomic sequence aligner, since a global (end-to-end) alignment strategy is doomed to frequently align unrelated regions, and worse, to produce misleading results for the common case of genome rearrangement, such as a family of duplicated genes. A related feature of PipMaker is that it can compare a finished sequence to a draft sequence and predict the orientation and ordering of the pieces having significant matches. (Although a prototype variant of PipMaker can compare two draft sequences and simultaneously order the pieces in each species, that capability is not currently available on the server. See Zhang *et al.*, 2001.) Another particular concern of PipMaker is to handle interspersed repeats in an appropriate manner: they are not permitted to align in the initial steps that determine the rough locations of matches, but can be aligned in later stages. This strategy avoids most spurious matches while permitting a repeat element that has assumed a functional role (e.g. Stavenhagen and Robins, 1988) to be detected if it occurs in both species.

PipMaker is designed for efficient comparison of two sequences of length about 100–1000 kb, and at an evolutionary distance approximately that of humans and mice. It is not designed to align just the regions with a conserved biological function; ideally it finds the orthologous nucleotide pairs, i.e. the position pairs that are descended from the same position in the ancestral sequence, allowing for substitution mutations. Under conditions that differ markedly from these assumptions, other methods may well be more appropriate.

For instance, an initial comparison of two entire chromosomes to identify homologous regions should be performed at higher speed and reduced sensitivity compared to PipMaker. An obvious and frequently effective approach is to find only gap-free alignments with very high scores, as sketched by Altschul *et al.* (1990, esp.

p. 409) and implemented by Schwartz *et al.* (1991). For extremely similar sequences there are 'greedy' alignment methods that compute optimal alignments. (Despite the name, in this context greedy methods are guaranteed to optimize an alignment score.) These algorithms allow gaps in the alignments and are extremely efficient, but work well only for very simple alignment-scoring schemes—for richer scores they lose their efficiency edge over dynamic programming. The basic techniques were developed by computer scientists in the mid-1980s, and have been useful for certain applications in bioinformatics (e.g. Florea *et al.*, 1998). Zhang *et al.* (2000) describe a variant that produces local alignments and survey the literature on this approach.

We believe that a more pressing need is for methods that give higher accuracy and/or more information than PipMaker offers. though perhaps at the cost of increased computation time. One potential approach is to stick with a dynamic programming alignment algorithm, but to use a more realistic scoring function. For instance, better approximations to the actual distribution of gap lengths (Gu and Li, 1995) can be used in optimal alignments, though at increased computational cost (Miller and Myers, 1988). Similarly, it is possible to score matches and mismatches in ways that may be more realistic (Huang, 1994). Another potentially useful approach for extracting better information by expending additional computational resources is through estimation of the reliability of each region within a computed alignment (Chao *et al.*, 1993b; Mevissen and Vingron, 1996; Holmes and Durbin, 1998).

A particularly attractive strategy is to apply hidden Markov models along the lines of Durbin *et al.* (1998, esp. Chapter 4), which can provide rigorous reliability estimations as well as segmenting the region based on degree of sequence conservation. Kent and Zahler (2000) describe an implementation of this approach. Another technique with considerable potential is a Gibbs sampling strategy (e.g. Wasserman *et al.*, 2000). Methods, such as these, with a rigorous statistical basis will be warmly received by biologists.

HOMOLOGY-ASSISTED GENE PREDICTION

As mentioned above, early genomic sequence alignments will be focused on finding protein-coding regions. Initial efforts to rigorously incorporate human–mouse comparisons into gene prediction methods have recently appeared (Bafna and Huson, 2000; Batzoglou *et al.*, 2000; Novichkov *et al.*, 2000; Wiehe *et al.*, 2000). However, some of these tools do not permit additional evidence concerning gene location to be utilized. There is a pressing need for a reliable tool that can accurately combine evidence from genomic sequence comparisons with the traditional clues from intrinsic sequence properties and the results of searching databases of protein sequences and ESTs.

MULTIPLE ALIGNMENT ALGORITHMS

The extensive literature on alignment methods for three or more sequences is almost entirely geared toward comparison of protein sequences. This is of course to be expected, since few examples exist of genomic sequence data from several similar species. However, that situation will change radically in the near future.

An early multiple alignment program aimed at genomic sequences is discussed by Hardison *et al.* (1994). It uses progressive alignment and quasi-natural gap costs (Altschul, 1989), which do about as well as possible at scoring gaps as dictated by a 'sum of pairwise scores' approach. Also, it pays considerable attention to effective utilization of computer space to obtain reasonable accuracy and efficiency. However, in our opinion, it requires too much control by the user. Our goal is a program that operates reliably in the absence of any user intervention. Such a program is part of our current prototype for MultiPipMaker, which will be released once we have an adequate variety of test data to warrant confidence in its reliability. However, suffice it to say that the problem is difficult, and that effort to improve upon existing solutions is appropriate.

VISUAL METAPHORS AND BROWSING TOOLS FOR ALIGNMENTS

The first people to contemplate genomic sequence alignments (e.g. Pustell and Kafatos, 1982) realized that visualization tools are necessary to cope with the potentially huge volume of output. Early work centered on the 'dotplot' representation (Schwartz *et al.*, 1991; Sonnhammer and Durbin, 1995), but there has been a shift of attention toward more compact representations (Chao *et al.*, 1993a; Koop and Hood, 1994; Duret *et al.*, 1996; Galili *et al.*, 1997; Jareborg and Durbin, 2000; Lund *et al.*, 2000; Göttgens *et al.*, 2001). An interesting variant of the problem is to effectively represent an alignment of two very similar sequences, which should emphasize the places where the sequences differ (Zhang and Madden, 1997; Delcher *et al.*, 1999).

More work is needed. For instance, the main visualization metaphor used by PipMaker, i.e. a 'percent identity plot' with one line per gap-free segment, is effective only for certain resolutions, say, 1–5 kb per inch of the figure; at lower resolution (i.e. more nucleotides per inch), it degenerates to a cloud of points conveying little if any information. What is the best way to summarize the varying degree of sequence conservation over a megabase region, using a picture that is, say, 1 inch high and 6 inches long? What about a summary for 500 bp in that same amount

Table 1. Some network resources for genomic sequence alignments. The following codes are used for Type: A = archived alignments, P = programs, and S = server

Name	http address	Type	Reference
Alfresco	http://www.sanger.ac.uk/Software/Alfresco	P	Jareborg and Durbin (2000)
CGAT	http://ftp.inertia.bs.jhmi.edu/roger/CGAT/CGAT.html	P	Lund *et al.* (2000)
EnteriX	http://ftp.globin.cse.psu.edu/enterix	A	Florea *et al.* (2000a)
GLASS	http://ftp.plover.lcs.mit.edu	S	Batzoglou *et al.* (2000)
Gibbs	http://www.wadsworth.org/res&res/bioinfo	P, S	Wasserman *et al.* (2000)
Intronerator	http://www.cse.ucsc.edu/~kent/intronerator	S	Kent and Zahler (2000)
LAJ	http://ftp.bio.cse.psu.edu	A, P	Wilson *et al.* (2001)
LAJ	http://ftp.web.uvic.ca/~bioweb/laj.html	A	Wilson *et al.* (2001)
MUMmer	http://www.tigr.org/softlab	P	Delcher *et al.* (1999)
PipMaker	http://ftp.bio.cse.psu.edu	S	Schwartz *et al.* (2000)
Rosetta	http://ftp.plover.lcs.mit.edu	S	Batzoglou *et al.* (2000)
SGP	http://ftp.soft.ice.mpg.de/sgp-1	S	Wiehe *et al.* (2000)
SynPlot	http://www.sanger.ac.uk/Users/jgrg/SynPlot	P	Göttgens *et al.* (2001)
VISTA	http://www.gsd.lbl.gov/vista	S	Dubchak *et al.* (2000)
WABA	http://www.cse.ucsc.edu/~kent/xenoAli/index.html	P, S	Kent and Zahler (2000)

of space? The difficult part of this is likely to lie in implementing an interactive software system that smoothly supports the chosen visual metaphors.

With multiple alignments, a few projects have explored issues of visualization and browsing (e.g. Schuler *et al.*, 1991; Boguski *et al.*, 1992; Jeanmougin *et al.*, 1998; Lee *et al.*, 1998; Dubchak *et al.*, 2000; Florea *et al.*, 2000a). Much more work along these lines will be appropriate and natural once a number of relevant datasets are in hand.

Potential components of alignment-browsing systems include tools to identify regions that exhibit properties suggestive of a particular biological function, such as matching the consensus sequence for a specific transcription factor binding site. Similarly, one might want tools that find particularly well conserved segments within an alignment (e.g. Stojanovic *et al.*, 1997, 1999).

Considerable impetus for further development of visualization/browsing techniques comes from the growing need for on-line archives of annotated alignments. A precomputed alignment, annotated with various kinds of hyperlinks, can present more detail than is possible in a traditional journal publication, and can be continuously updated. Internet archives of genomic alignments exist for *E.coli* and several closely related organisms (EnteriX; Florea *et al.*, 2000a), and for *C.elegans* and a close relative (Intronerator; Kent and Zahler, 2000). Other sites give a preview of how this might work for mammalian genomes (LAJ; Wilson *et al.*, 2001).

EVALUATING ALIGNMENT METHODS

There is an urgent need for methods to evaluate the effectiveness of alignment software for genomic sequences. The situation stands in stark contrast to that for software that aligns protein sequences, where there exist well-curated datasets of 'correct' alignments and established protocols for their use in software evaluation (e.g. Thompson *et al.*, 1999).

Of course, for alignment tools intended for gene prediction, one has the benefit of an extensive literature and several large datasets for evaluating *ab initio* methods. Instead, the problem lies with evaluating methods aimed largely at properly aligning non-coding regions, since we rarely know what the 'right answer' is. At first glance, the problem seems tractable—we can extract examples from some available database of experimentally confirmed regulatory sites, such as TRRD (Kolchanov *et al.*, 2000), and measure each program's ability to detect those regions. However, in our hands (Stojanovic *et al.*, 1999; Florea *et al.*, 2000b) such an approach proved to be far more difficult than initially imagined. A prime example of software evaluation in this area is given by Wasserman *et al.* (2000).

URLS

Table 1 collects together the World-Wide Web addresses of some of the tools discussed above.

DISCUSSION

The lure of effectively utilizing the forthcoming bounty of genome sequence data from two or more closely related organisms will trigger an explosion of new ideas and software. Now is the time for an unfettered exploration of the possibilities by all interested parties. Our aim in writing this report is to assist individuals and groups wanting to quickly familiarize themselves with this exciting and promising area. The near-term research directions identi-

fied above need to be addressed as soon as possible, which will require the combined ideas and concerted effort of many enthusiastic researchers.

However, there is another compelling need that should not be overlooked. The bioinformatics community will be ill-served if we produce a bewildering array of tools with capabilities and strengths that overlap in a complex manner. Of course, it might turn out that one individual or small group is able to produce a product of such high quality that a biologist can safely forget the other tools; for mammalian genome sequences this happened with *ab initio* gene prediction and with identification of interspersed repeats. Given the extreme time pressures and the wide range of expertise required, we doubt that any one group will soon solve all of the problems outlined here.

We all know of areas within bioinformatics where there are numerous software tools, none of which is clearly superior. Unfortunately for users, an area can become quite cluttered with mediocre tools, because many people find it much easier and more fun to develop a new program than to adequately verify that it actually improves upon earlier work, and using another person's software is sometimes treated like using their toothbrush. The introduction of new programs without a careful evaluation can run counter to the interests of the biomedical community. Individual biologists may waste time doing their own, often incomplete, comparisons among a larger collection of competing tools or, worse yet, a trusting user may draw conclusions from improper or inferior output. We are already seeing signs of spurious clutter in some of the areas surveyed above.

However, within a few years it will be technically feasible for the bioinformaticians with expertise in developing software for comparing genomic DNA sequences to pool their ideas and energy to produce a compact tool set that serves a number of needs of biomedical researchers. We hope that the individuals involved will be sufficiently community-minded to do so.

ACKNOWLEDGEMENTS

This work was supported by grant HG02238 from the National Human Genome Research Institute.

REFERENCES

Altschul,S. (1989) Gap costs for multiple sequence alignment. *J. Theor. Biol.*, **138**, 297–309.

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST—a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ansari-Lari,M.A., Oeltjen,J.C., Schwartz,S., Zheng,Z., Muzny,D.M., Lu,J., Gorrell,J.H., Chinault,A.C., Belmont,J.W., Miller,W. and Gibbs,R.A. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region on mouse chromosome 6. *Genome Res.*, **8**, 29–40.

Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. *Ismb*, **8**, 3–12.

Batzoglou,S., Pachter,L., Mesirov,J., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.

Boguski,M., Hardison,R., Schwartz,S. and Miller,W. (1992) Analysis of conserved domains and sequence motifs in cellular regulatory proteins and locus control regions using new software tools for multiple alignment and visualization. *The New Biologist*, **4**, 247–260.

Chao,K.-M., Hardison,R. and Miller,W. (1993a) Constrained sequence alignment. *Bull. Math. Biol.*, **55**, 503–524.

Chao,K.-M., Hardison,R. and Miller,W. (1993b) Locating well-conserved regions within a pairwise alignment. *CABIOS*, **9**, 387–396.

Delcher,A.L., Kasif,S., Fleischman,R., Peterson,J., White,O. and Salzberg,S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27**, 2369–2376.

Dubchak,I., Brudno,M., Loots,G., Pachter,L., Mayor,C., Rubin,E. and Frazer,K.A. (2000) Active conservation of noncoding sequences revealed by three-way species comparison. *Genome Res.*, **10**, 1304–1306.

Dunham,I., Shimizu,N., Roe,B., Chissoe,S. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Duret,L., Gasteiger,E. and Perriere,G. (1996) LALNVIEW: a graphical viewer for pairwise sequence alignments. *CABIOS*, **12**, 507–510.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Florea,L., Riemer,C., Schwartz,S., Zhang,Z., Stojanovic,N., Miller,W. and McClelland,M. (2000a) Web-based visualization tools for bacterial genome alignments. *Nucleic Acids Res.*, **28**, 3486–3496.

Florea,L., Li,M., Riemer,C., Giardine,B., Miller,W. and Hardison,R. (2000b) Validating computer programs for functional genomics in gene regulatory regions. *Current Genomics*, **1**, 11–27.

Galili,N., Baldwin,H., Lund,J., Reeves,R., Gong,W., Wang,Z., Roe,B., Emanuel,B., Nayak,S., Mickanin,C., Budarf,M. and Buck,C.A. (1997) A region of mouse chromosome 16 is syntenic to the DiGeorge, velocardiofacial syndrome minimal critical region. *Genome Res.*, **7**, 17–26.

Göttgens,B., Gilbert,J., Barton,L., Grafham,D., Rodgers,J., Bentley,D. and Green,A.R. (2001) Long range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved non-coding sequences. *Genome Res.*, **11**, 87–97.

Gu,X. and Li,W.-H. (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignments. *J. Mol. Evol.*, **40**, 464–473.

Guigó,R., Agarwal,P., Abir,J., Burset,M. and Fickett,J.W. (2000) An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, **10**, 1630–1642.

Hardison,R. and Miller,W. (1993) Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.*, **10**, 73–102.

Hardison,R., Chao,K.-M., Schwartz,S., Stojanovic,N., Ganetsky,M. and Miller,W. (1994) Globin Gene Server: a prototype E-mail database server featuring extensive multiple alignments and data compilation for electronic genetic analysis. *Genomics*, **21**, 344–353.

Hardison,R., Oeltjen,J. and Miller,W. (1997) Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7**, 959–966.

Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *J. Comput. Biol.*, **5**, 493–504.

Huang,X. (1994) A context dependent method for comparing sequences. *Proceedings of the 5th Symposium on Combinatorial Pattern Matching Lecture Notes in Computer Science 807*, Springer, Berlin, pp. 54–63.

Jang,W., Hua,A., Spilson,S.V., Miller,W., Roe,B.A. and Meisler,M.H. (1999) Comparative sequence of human and mouse BAC clones from the *mnd2* region of chromosome 2p13. *Genome Res.*, **9**, 53–61.

Jareborg,N. and Durbin,R. (2000) Alfresco—a workbench for comparative genomic sequence analysis. *Genome Res.*, **10**, 1148–1157.

Jareborg,N., Birney,E. and Durbin,R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.*, **9**, 815–824.

Jeanmougin,F., Thompson,J., Gouy,M., Higgins,D. and Gibson,T.J. (1998) Multiple sequence alignment with ClustalX. *Trends Biol. Sci.*, **23**, 403–405.

Kent,W.J. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale *C.briggsae–C.elegans* genomic alignment. *Genome Res.*, **10**, 1115–1125.

Kolchanov,N.A., Podkolodnaya,O., Ananko,E., Ignatieva,E., Stepanenko,I., Kel-Margoulis,O., Kel,A., Merkulova,T., Goryachkovskaya,T., Busygina,T., Kolpakov,F., Podkolodny,N., Naumochkin,A., Korostishevskaya,I., Romashchenko,A. and Overton,G.C. (2000) Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Res.*, **28**, 298–301.

Koop,B.F. and Hood,L. (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genet.*, **7**, 48–53.

Lee,I.Y., Westaway,D., Smit,A., Wang,K., Seto,J., Chen,L., Acharya,C., Ankener,M., Baskin,D., Cooper,C., Yao,H., Prusiner,S. and Hood,L.E. (1998) Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. *Genome Res.*, **8**, 1022–1037.

Lund,J., Chen,F., Hua,A., Roe,B., Budarf,M., Emanuel,B. and Reeves,R.H. (2000) 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*, **63**, 374–383.

McClelland,M., Florea,L., Sanderson,K., Clifton,S., Parkhill,J., Churcher,C., Dougan,G., Wilson,R. and Miller,W. (2000) Comparison of the *Escherichia coli* K12 genome with sampled genomes of *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res.*, **28**, 4974–4986.

Makalowski,W., Zhang,J. and Boguski,M. (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.*, **6**, 846–857.

Mevissen,H.T. and Vingron,M. (1996) Quantifying the local reliability of a sequence alignment. *Protein Eng.*, **9**, 127–132.

Miller,W. and Myers,E. (1988) Sequence comparison with concave weighting functions. *Bull. Math. Biol.*, **50**, 97–120.

Morgenstern,B., Frech,K., Dress,A. and Werner,T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.

Novichkov,P., Gel'fand,M. and Mironov,A. (2000) Prediction of the exon–intron structure by comparison of nucleotide sequences from various genomes. *Mol. Biol.*, **34**, 230–236. (In Russian).

Onyango,P., Miller,W., Lehoczky,J., Leung,C., Birren,B., Wheelan,S., Dewar,K. and Feinberg,A.P. (2000) Sequence and comparative analysis of the mouse 1 megabase region orthologous to the human 11p15 imprinted domain. *Genome Res.*, **10**, 1697–1710.

Pustell,J. and Kafatos,F.C. (1982) A high speed, high capacity homology matrix: zooming through SV40 and polyoma. *Nucleic Acids Res.*, **10**, 4765–4782.

Schuler,G.D., Altschul,S. and Lipman,D. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.

Schwartz,S., Miller,W., Yang,C.-M. and Hardison,R. (1991) Software tools for analyzing pairwise alignments of long sequences. *Nucleic Acids Res.*, **19**, 4663–4667.

Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular sequences. *J. Mol. Biol.*, **147**, 195–197.

Sonnhammer,E.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1-10.

Stavenhagen,J.B. and Robins,D.M. (1988) An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene. *Cell*, **55**, 247–254.

Stojanovic,N., Berman,P., Gumucio,D., Hardison,R. and Miller,W. (1997) A linear-time algorithm for the 1-mismatch problem. In Dehne,F., Rau-Chaplin,A., Sack,J.-R. and Tamassia,R. (eds), *Algorithms and Data Structures Lecture Notes in Computer Science, 1272*, Springer, Berlin, pp. 126–135.

Stojanovic,N., Florea,L., Riemer,C., Gumucio,D., Slightom,J., Goodman,M., Miller,W. and Hardison,R. (1999) Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.*, **27**, 3899–3910.

Thompson,J.D., Plewniak,F. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Wasserman,W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.

Wiehe,T., Guigó,R. and Miller,W. (2000) Genome sequence comparisons: hurdles in the fast lane to functional genomics. *Briefings in Bioinformatics*, **1**, 381–388.

Wilson,M.D., Riemer,C., Martindale,D., Schnupf,P., Boright,A., Cheung,T., Hardy,D., Schwartz,S., Scherer,S., Tsui,L.-C., Miller,W. and Koop,B.F. (2001) Comparative analysis of the gene dense *ACHE/TFR2* region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res.*, **29**, 1352–1365.

Zhang,J. and Madden,T.L. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.*, **7**, 649–656.

Zhang,Z., Berman,P. and Miller,W. (1998) Alignments without low-scoring regions. *J. Comput. Biol.*, **5**, 197–210.

Zhang,Z., Berman,P., Wiehe,T. and Miller,W. (1999) Post-processing long pairwise alignments. *Bioinformatics*, **15**, 1012–1019.

Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.

Zhang,Z., Berman,P., Schwartz,S., Bouck,J. and Miller,W. (2001) Aligning two fragmented genomic sequences. *Bioinformatics*, in press.