



Tricross: using dot-plots in sequence-id space to detect uncataloged intergenic features

William C. Ray^{1, 2,*}, Robert S. Munson Jr^{1, 2, 3} and Charles J. Daniels³

¹Children's Research Institute, ²The Department of Pediatrics, The Ohio State University, 700 Childrens Dr Columbus, OH 43205, USA and ³The Department of Microbiology, The Ohio State University, 484 West 12th Ave, Columbus, OH 43210, USA

Received on April 13, 2001; revised on June 15, 2001; accepted on June 26, 2001

ABSTRACT

Motivation: The process of determining the functional sequence content of an organism is confounded by several factors. Large protein coding sequences are relatively easy to find by statistical methods. Smaller proteins however may escape detection due to their size falling below some arbitrary researcher-defined minimum cutoff, or the inability to precisely define a promoter, or translational start (Delcher *et al.*, *Nucleic Acids Res.*, **27**, 4636–4641, 1999). Promoter and regulatory sequences themselves are difficult to define due to a significant amount of allowable sequence variation, as well as a probable lack of any completely accurate whole-organismal gene catalogs to date. Finally, certain genes coding functional RNAs may have insufficient structural or sequence constraints to be detectable by normal sequence structure/pattern searching methods (Eddy and Rivas, *Bioinformatics*, **16**, 583–605, 2000).

In those cases where there are multiple closely related organisms that have been sequenced, there is additional information that may be used in the investigation of sequence content—that being the possible conserved nature of functional sequences between the organisms.

We present a method for the utilization of this conserved information to detect genes and other potentially functional sequences that may be missed by standard ORF-calling, RNA finding, and pattern matching software. The *tricross* programs produce a multi-way cross comparison of three sets of sequences, determine which are conserved in all three sets, and produce a graphical (Virtual Reality Modelling Language—VRML; (ISO/IEC 14 772-1: 1997, VDC), 1997) representation as well as alignments of all sequence triples found. The software can also be applied to a pair of sequence sets, though the noise in the results increases.

Results: *Tricross* has been used to examine the intergenic-sequence content of the three archaeal *Pyrococcus* genomes to determine the most highly related sequences remaining between the annotated protein and RNA coding sequences. Set to relatively stringent similarity requirements for the search, *tricross* found 101 intergenic sequences conserved among the three organisms. Interestingly, 29 of these appear to contain members of a family of small RNA molecules (Kiss-Laszlo *et al.*, *EMBO J.*, **17**, 797–807, 1998) only recently discovered in the Archaea (Armbruster, OSU, Diss., 1988; Omer *et al.*, *Science*, **288**, 517–522, 2000; Gaspin *et al.*, *J. Mol. Biol.*, **297**, 895–906, 2000). While some of the remaining 72 appear to be individual highly conserved promoter sequences, others have no currently known biological significance.

Although originally developed to facilitate the examination of intergenic sequences, none of the *tricross* logic is inherently specific to intergenic sequences. The software can also be applied to gene sequences, and has been used to produce inter-genomic gene order dot-plots for *Haemophilus influenzae* (Fleischmann *et al.*, *Science*, **269**, 496–512, 1995) versus *H. ducreyi* (unpublished data), and *Neisseria meningitidis* Z2491 (serogroup A) (Parkhill *et al.*, *Nature*, **404**, 502–506, 2000) versus *Neisseria meningitidis* Z58 (serogroup B) (Tettelin *et al.*, *Science*, **287**, 1809–1815, 2000) versus *Neisseria gonorrhoeae* (Lewis *et al.*, <http://micro-gen.ouhsc.edu/>, 2000).

Availability: The *tricross* software package is available from <http://www.biosci.ohio-state.edu/~ray/bioinformatics/tricross.html>.

Contact: ray@biosci.ohio-state.edu; daniels.7@osu.edu; munsonr@pediatrics.ohio-state.edu

Supplementary information: Additional data from the cross-genomic comparisons examined in the discussion section are linked from <http://www.biosci.ohio-state.edu/~ray/bioinformatics/tricross.html>.

*To whom correspondence should be addressed.

INTRODUCTION

The examination of sequences between known genes is interesting on several levels.

At the most obvious, areas between known genes are a logical place to search for as-yet unknown genes. ORF-calling, the process of predicting the location and bounds of probable protein coding regions, is an inexact science and is likely to remain so in the future. The sequence criteria regarding what constitutes a probable protein coding region are sufficiently variable that researchers searching for coding regions must make an arbitrary decision regarding the smallest acceptable coding region. This decision trades reduction in false calls against the potential to overlook small transcribed and translated sequences. A similar problem arises from the difficulty in identification of the correct start codon. The point, in fact, is that the process is probabilistic, and is unlikely to ever be reduced to a zero error rate.

Additional interest in the regions between known genes stems from the fact that these are areas likely to contain promoter or regulatory motifs. For example, the discovery that the archaeal transcription machinery closely resembles Pol II transcription in eukaryotes, and the early availability of multiple archaeal genomic sequences, made the investigation of archaeal transcription a simplified window on eukaryal transcription. Information regarding similarities discovered in promoter regions may also be useful for determining coordinately regulated gene families, or other regulatory motifs.

Finally, ORF-calling is concerned only with the discovery of protein-coding genes, and does not apply to RNA genes, which do not require the existence of an open reading frame. Existing RNA gene finding programs rely on searching for sequences with high similarity to a known sequence and/or a known structure. Eukaryotes, and more recently Archae have been discovered to have a number of small RNAs that have insufficient structural or sequence similarity to be found using sequence and structural homology alone. These 'small nucleolar RNAs' or snoRNAs, are a known part of a modification scheme for rRNAs (Kiss-Laszlo *et al.*, 1998) and have been implicated in participation in the same modification of tRNAs (Armbruster, 1998; Omer *et al.*, 2000). Due to the presence of only a small required sequence motif, they can only be uniquely identified in a sequence searching context by use of knowledge about the target site for the structural RNA modification (Eddy and Rivas, 2000).

Early attempts at examination of archaeal intergenic sequences for cross-genomic conservation demonstrated that naive sequence-similarity based searches of all intergenic sequences against all intergenic sequences produced intractable results. Not only did the search produce, for *Methanobacterium thermoautotrophicum* versus *Archaeoglobus fulgidis*, 70 Mb of data, but the most

significant hits discovered were between 'lost' portions of known genes appearing in the intergenic regions due to misannotation of start sites.

The publication of the three archaeal *Pyrococcus* strains (*Pyrococcus horikoshii*; Kawarabayasi *et al.*, 1998, *Pyrococcus abyssi*; Heilig and Genoscope, 1999, and *Pyrococcus furiosus*; Utah, 1999), and the development of a method to reduce the quantity of erroneously included sequences (Ray and Daniels, 2001), provided sufficient data to try a revised approach to the problem.

Embodied in the *tricross* programs, our revised approach is essentially a modification and repurposing of the well-known dot-plot method of cross-sequence comparison. Dot-plots produced by *tricross* are in sequence-identifier, rather than sequence-coordinate space, and reflect the degree of mutual conservation of a sequence rather than the degree of similarity in the conservation.

Utilizing this methodology, we have examined the intergenic regions of the *Pyrococcus* genomes and discovered an interesting collection of sequences. The applicability of sequence-identifier based dot-plots extends outside the realm of conserved sequence discovery as well, and has proven a natural method for the examination of global genome ordering in the multiple sequenced *Haemophilus* and *Neisseria* genomes.

SYSTEM AND METHODS

The *tricross* programs are implemented as a csh shell script, and a pair of Perl programs. They additionally make use of the FASTA (Pearson and Lipman, 1988) sequence search program, and the Clustal-W (Thompson *et al.*, 1994) sequence alignment program. The *tricross* programs have been successfully executed on a SunOS 4.1.4 platform, a Solaris 7 platform, and a Digital Unix 4.0 platform without modification. Two slightly different implementations of the algorithm allow the user to trade memory for execution speed, and in the more memory-intensive version, roughly 200 Mb main memory is required for a three-dimensional comparison of three sets of 2500 sequences each.

The results provided in the discussion are based on FASTA version 2.0x run on an UltraSparc Enterprise 250 running Solaris 7. The FASTA version used, and/or system math libraries, appear to affect the calculated FASTA score to some degree. Using a different platform or different FASTA version results in slight variations in the number of sequences that meet the cutoff criteria.

ALGORITHM

Normally a dot-plot is constructed by examining percentage similarity in a pair of sliding windows in the sequences being compared. A score is calculated based on the per-

centage similarity, and a two-dimensional plot constructed with the set of windows along the sequences on each axis. The window–window scores are plotted strictly by cutoff, or by some extra-dimensional visualization technique such as color temperature.

To repurpose dot-plot methodology for the question at hand, we start with the realization that the basic concept of a dot-plot as a similarity finder between a pair of sequences, can also be applied as a dissimilarity rejector by the addition of a third sequence and the requirement that a match must exist in it also, for a found ‘dot’ to be significant. This allows a dot-plot in three dimensions to more efficiently reject the sort of noise matches that corrupted our early naive cross-genomic comparisons. For a lost gene fragment to appear as a match in a three-dimensional dot-plot, the start must be similarly misannotated in not just two, but all three organisms.

In addition, we replace the sliding window similarity detection method of a traditional dot-plot with a per-sequence similarity calculation, in our implementation performed by FASTA. The dot-plot therefore is performed between ordered lists of identified sequences, and the results are plotted in sequence-identifier space, rather than in sequence-coordinate space. Dots in this model represent, in two dimensions, pairs of sequences that are ‘best neighbors’ (most closely related sequences) between two sets of sequences, and in three dimensions, best neighbor triples.

Finally, we are less interested in the similarity score with which a pair or triple of sequences are found to be best neighbors, than in applying a rating of the neighborliness of the relationship. For this purpose we introduce the concept of the ‘N-way’ness of the relationship, which is a measure of the mutuality with which the sequences find each other as best neighbors.

Itemized, the search with three genomes (collections of sequences) is as follows, with each of the enumerated blocks embodied in separate programs that successively handle the data:

- (1) For each sequence in a genome, find its best match (or best matches) in each other genome. Record results as a list of neighbors for each gene. Repeat for all genomes to be compared.
- (2) For each potential sequence triple, check neighbor lists for each member gene to determine the N-way hit relationship between the genes in the candidate triple. If the N-way relationship is large enough, output triple and hit information into a table.
- (3) Filter triples matching minimum N-way hit requirements from step 2, for hits that exceed a user-specified display-time cutoff, and write the corresponding data into a Virtual Reality Modelling

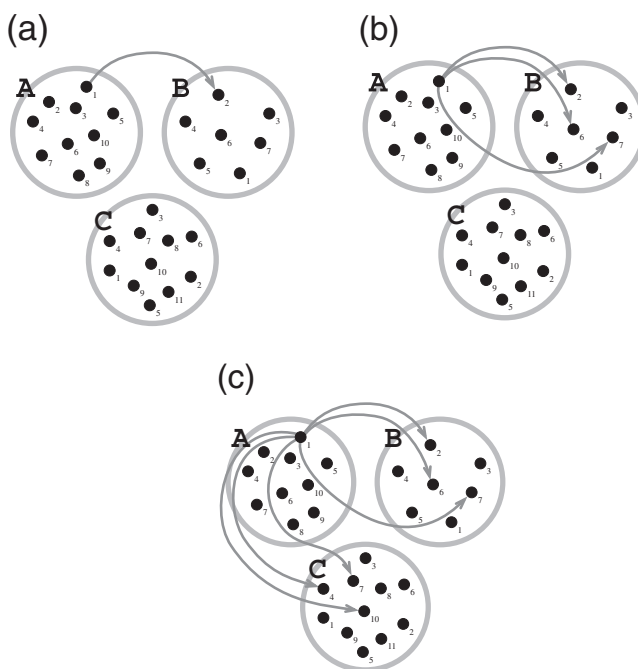


Fig. 1. For each sequence in each genome, FASTA is used to find its closest relatives in the other genomes. The number of relatives accepted from FASTA can be changed, and setting it to single best relatives, or the top n relatives produces differing data. (a) For each sequence that belongs to genome A, FASTA is used to find its closest relative in genome B. A directed edge from sequence i in genome A to sequence j in genome B indicates that a FASTA search using sequence i as the query finds sequence j as a relative. (b) Depending on the number of FASTA hits accepted, there may be a number of hits from i to sequences in genome B. (c) Both genome B and C are searched for neighbors of each sequence i in A. This process is then repeated for each sequence j in B, and each sequence k in C.

Language (VRML) file. Link sequence triples in the 3D VRML file to alignments of the sequences.

Schematically, the comparisons are shown and described in Figures 1 and 2.

While the comparisons shown are for a small number of sequences, one can think of the total genome–genome–genome comparison as the creation of a large directed graph. Sequences become graph nodes, and the existence of a FASTA-determined hit from one sequence to another places a directed edge between these two sequences. Depending on the intent, different information can be discovered by restricting the FASTA hits to only the singular best hit in each of the other organisms, or to some arbitrary number or quality of the top hits.

The degree of the mutual relationship between sequences is denoted in terms of the ‘N-way’ness of the relationship. For any given triple of sequences, this degree is the number of edges between the three sequences in

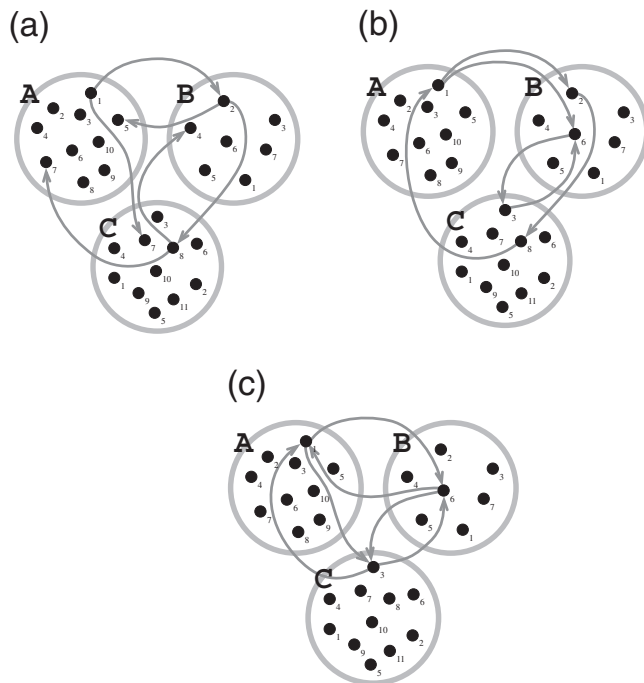


Fig. 2. Once the comparisons shown in Figure 1 are computed, sequence triples can be ranked based on how many edges they share between them. (a) Once all the edges in the digraph have been created, they can be examined for relationships among the sequences. While FASTA has found sequence neighbors in the diagram shown, the significance is not apparent. (b) Here, sequences have found neighbors that find them again in return. This is an example of a ‘3-way’ hit between the triple (1, 2, 8) in genomes A–C respectively. (1, 6, 3) likewise displays a 3-way hit. (c) In this figure, the triple (1, 6, 3) displays a 6-way hit. If more than one sequence neighbor is accepted from FASTA, any given sequence may participate in a number of triples of varying degrees with other sequences in the other genomes.

the triple in the directed graph generated by the FASTA searches.

A 6-way hit between a triple of sequences indicates that for each sequence in the triple, the most similar sequence in each of the other genomes also finds this sequence as its closest relative.

A 3-way hit between a triple of sequences is the minimum required to ensure that FASTA in fact found a relationship between all three sequences. This degree of relatedness may be indicative of a single cycle between the three sequences, but also occurs whenever one sequence’s hit in either of the other genomes finds that sequence again.

If the FASTA searches are restricted to only the singular best hit in the two other organisms, then for each gene in an organism, in the digraph only two edges leave that gene, travelling to one gene in each of the other organisms. A

6-way hit for a triple indicates that all the members of the triple are completely connected in the digraph—no edges leave to other genes outside the triple. In other words each finds the other members of the triple as their most highly related sequences in the other genomes. These sequences represent what appear to be unique conserved sequence regions occurring between coding regions.

If the FASTA searches are allowed to return multiple top-hits for a given sequence, the result is that the digraph now has the potential that for a sequence in genome ‘A’, multiple edges may leave to sequences in genome ‘B’. In this case, a 6-way hit for a triple indicates complete, but not exclusive connection among the members. In other words each finds the other members as a highly related sequence, but not necessarily the most highly related. These sequences represent what may be families of conserved sequences or conserved sequence motifs occurring between coding regions.

Although considering 5, 4 or 3-way hits results in the inclusion of additional sequences with decreasing requirements for mutual relationship, which decreases the likelihood of finding unique shared sequence regions, these hits increase the information available about patterns in the relationships between the genomes.

IMPLEMENTATION

Input and output

Tricross takes as input three multi-sequence FASTA files. Each sequence in these files is required to have a unique name for the sequence, positioned as the first word following the FASTA header ‘>’. *Tricross* also requires a set of three files defining the ordering of the sequences in sequence-id space.

Tricross produces as output a collection of FASTA search files, one for each sequence, searched into each genome. The number of FASTA hits accepted from each search, and the FASTA score cutoff that the software should consider to be significant is configurable in the software. It also produces a list of sequence triples sharing N or more hits, with N being a user-configurable parameter. For each triple of sequences in the list, it produces a Clustal-W alignment. Finally, it produces a VRML file containing a representation of the discovered sequence relationships in three-dimensional sequence-id space. In this visualization file, the plotted representative for each triple is linked to the Clustal-W alignment file for that triple.

A pair of sequence sets, such as the collections of proteins for two genomes, can also be compared, though obviously this limits the relationships to a 2-way maximum. A separate implementation of *tricross* is provided to handle two-way comparisons, though the algorithm is identical to that discussed for three-way comparisons, with the third

sequence collection considered to be empty.

Limitations

When implemented to execute as described, there is nothing inherent to the logic that precludes the inclusion of more than three sequence sets in the analysis. While the lack of a simple way to present N -dimensional data, with $N > 3$ is a practical limitation of the visualization component, the list file and alignments would still be of use. Unfortunately, the execution time is a larger practical limitation on increasing the number of sequence sets under consideration. Step 2 of the algorithm as presented, examination of all possible sequence triples for those with high N -way hit scores, requires an average of $O(\binom{k}{2}n)$ memory, and $O(n^k)$ time for k collections of sequences with n sequences in each.

For three genomes, the algorithm described can be modified to eliminate the $O(n^3)$ time requirement of examining n^3 triples, by summing hits directly into the elements of an $n \times n \times n$ matrix. It might initially appear that this would increase the memory cost to $O(n^3)$, but useful cross-genomic comparisons may be expected to be characterized by each sequence finding a small number of neighbor sequences compared to the total number of candidate sequences. Because of this, the filled locations in the $n \times n \times n$ matrix are sparse, and the use of a sparse matrix storage method allows the average memory cost to increase only to $O(n^2)$, while the execution time falls to $O(n^2)^\dagger$.

Currently, the $O(n^3)$ time implementation is capable of comparing the three *Pyrococcus* genomes' intergenic regions (some 1200 sequences each) overnight when executed on a 250 Mhz Sun Ultrasparc platform. The $O(n^2)$ implementation is capable of comparing the three *Neisseria* genomes' coding sequence regions (average of 2500 sequences each) in 4 h on a 500 Mhz DEC Alpha platform.

DISCUSSION

When the high confidence intergenic regions are extracted (using the PACRAT system; Ray and Daniels, 2001) from *P.furiosus*, *P.horikoshii*, and *P.abysyi*, the result is three multiple-sequence FASTA files containing 1291, 1199, and 1144 sequences respectively. With the *tricross* FASTA search constrained to return only the single best match with EScore better than 1×10^{-3} , the intergenomic similarities found are as shown in Table 1.

Amongst these 6 sets of genome–genome hits, 103 are found to be 6-way hits, and 31 are found to be 5-way hits among all three genomes. Additional examination reveals that 33 of the sequence triples have been found twice, due

Table 1. The number of similar sequences found between each pair of *Pyrococcus* genomes' intergenic sequences, allowing a single top FASTA hit with score better than $E = 0.001$

From genome	To genome	Number of FASTA similarities to better than $E = 0.001$
<i>P.abysyi</i>	<i>P.furiosus</i>	234
<i>P.abysyi</i>	<i>P.horikoshii</i>	414
<i>P.horikoshii</i>	<i>P.furiosus</i>	240
<i>P.horikoshii</i>	<i>P.abysyi</i>	432
<i>P.furiosus</i>	<i>P.horikoshii</i>	248
<i>P.furiosus</i>	<i>P.abysyi</i>	242

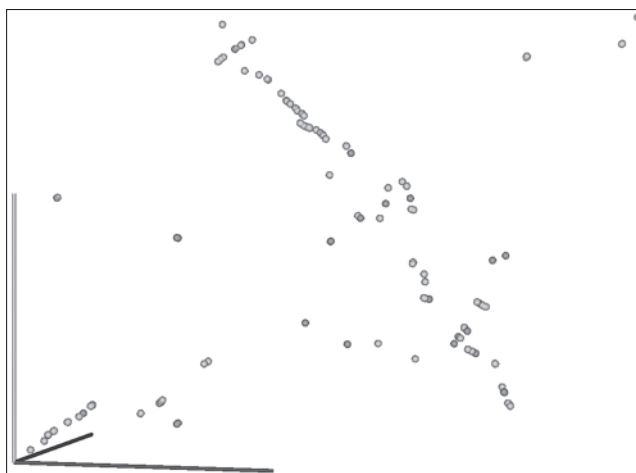


Fig. 3. An off-axis view of the 3-way genome comparison VRML world produced by *tricross* run on the three *Pyrococcus* genomes. *P.horikoshii* is represented on the vertical y -axis, *P.abysyi* is represented on the z -axis which proceeds left to right in this depiction, and *P.furiosus* is represented along the x -axis which is receding. 101 spheres appear in this view, representing sequence triples with 5-way or better cross-genomic relationships from FASTA singular top hits.

to some sequences being included in the search set as the upstream relative of each of two divergent genes, or the downstream relative of two convergent genes. Remaining are 101 unique sequence triples that have scored either 6-way, or 5-way unique mutual relationships among the members of the triple.

Figure 3 shows a snapshot of the genome-space depicting the relationships, with *P.furiosus* represented along the x -axis (receding), *P.horikoshii* along the y -axis (vertical) and *P.abysyi* along the z -axis.

Looking at the aligned sequences returned by this constrained run of the *tricross* programs, it appears that many of the sequences returned show familial resemblances. Unlike the results of naive genome–genome FASTA comparisons, the data from this analysis is approachable

[†] No claim is made that optimal data structures or algorithms have been used in this implementation.

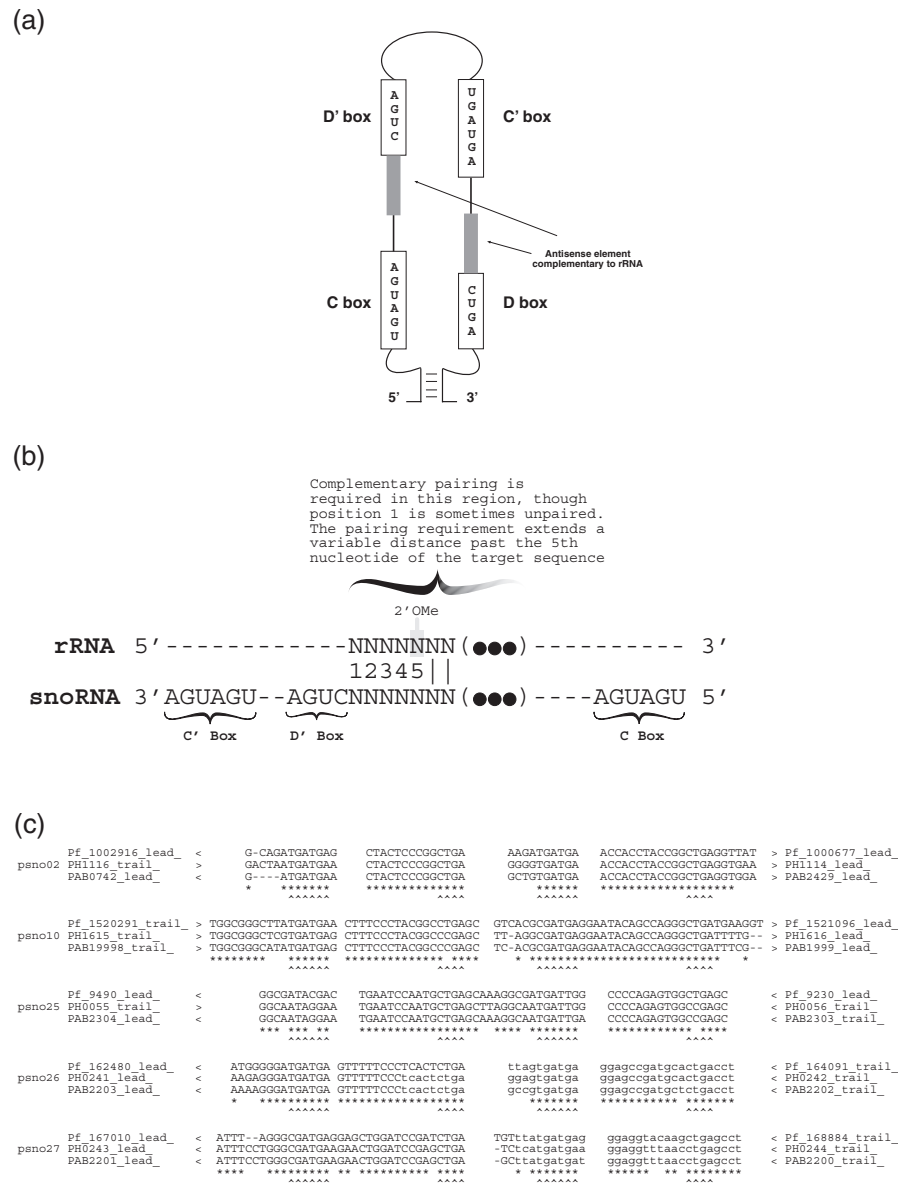


Fig. 4. Typical snoRNA sequence motifs, and several representative examples found in the Archaeal *Pyrococcus* genomes. (a) A schematic diagram of the snoRNA consensus. Some snoRNAs lack the *D'Box* and the *C'Box* shown in this figure, having only one single target sequence. Many snoRNAs have regions which can pair to bring the ends of the snoRNA together, located outside the *CBox* and *DBox*. (b) The snoRNA pairs with a stable RNA by standard Watson-Crick pairing of the snoRNA targeting region sequence to the stable RNA molecule. At the 5th position from the boundary of the *DBox* or *D'Box*, a methyl group is attached to the stable ribose 2' oxygen. (c) Aligned snoRNA-like patterns discovered by a search for mutually conserved sequences in the three sequenced *Pyrococcus* genomes. The *C* and *DBox* motifs are aligned for each sequence, and denoted with ~~~~ marks under their locations. The ORFs upstream and downstream of the snoRNA pattern are annotated at the borders of the table, as well as the intergenic region's relationship to the ORF, and the ORF's orientation with respect to the snoRNA. Positions shown in lower-case letters are not intergenic, but from locations where the snoRNA-like pattern extends into the bounding ORF, and are shown to demonstrate that these sequences are continuous and consistent with the patterns found entirely within the intergenic sequences.

visually. Inspection of the collected alignments leads to the conclusion that there are several recurring sequence motifs into which the majority of the sequences fall. A complete presentation of the aligned sequences, being too large to place in this paper, is available from <http://www.biosci.ohio-state.edu/~ray/bioinformatics/tricross.html>.

Of this collection however, one of the most obvious patterns, and the most straightforward to ascertain the potential biological relevance of, is one that resembles that of a class of snoRNAs known as *C/Dbox* snoRNAs.

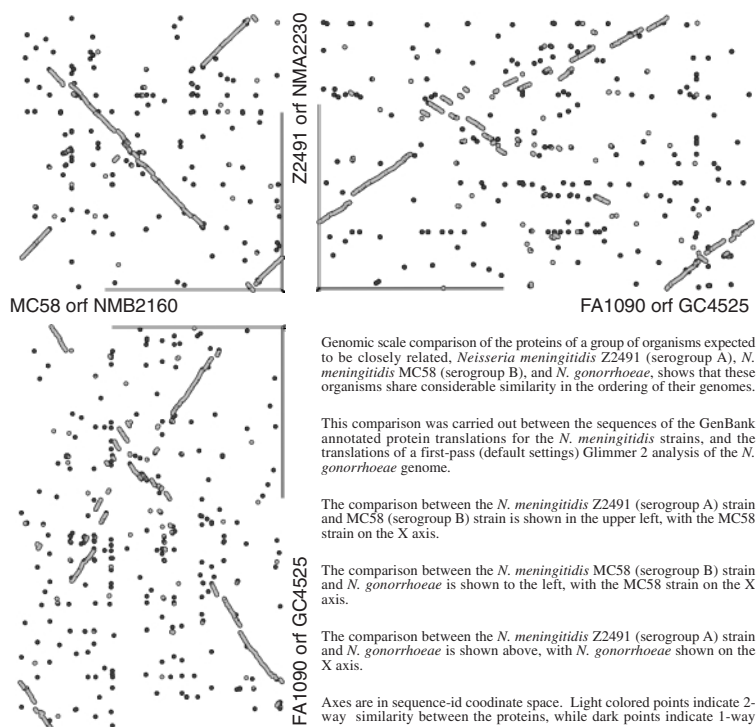


Fig. 5. Three 2-way comparisons between the collected protein sequences of the three sequenced *Neisseria* genomes.

A schematic diagram of *C/Dbox* snoRNAs is shown in Figure 4a, while the snoRNA's participation in targeting the stable-RNA modification is shown in Figure 4b. (For a complete discussion of snoRNAs and their participation in the modification system in Archaea, see Armbruster, 1998; Gaspin *et al.*, 2000, or Kiss-Laszlo *et al.*, 1998). Figure 4c shows several interesting members of the subset of sequence triples that appear to be *C/Dbox* snoRNAs. Again, a full presentation of the data is precluded here by space limitations, and is available from <http://www.biosci.ohio-state.edu/~ray/bioinformatics/tricross.html>.

While effective in detecting highly conserved sequences, *tricross* detects only 30 potential snoRNAs in the *Pyrococcus* genomes (28 intergenic sequence triples contain what appears to be single snoRNA genes, 1 triple contains a divergent pair of potential snoRNA genes). This is significantly less than 46 (Gaspin *et al.*, 2000) or more (Omer *et al.*, 2000) predicted by searches specifically directed at finding snoRNAs in Archaea. This discrepancy however is expected.

First, the sequences examined which produced the results presented here were pre-screened to remove potential gene fragments. As *tricross* requires that the sequence be found in all three genomes, a mis-called gene start in one genome can mask one member of a sequence triple with similarity, and prevent its discovery. Since only approximately 60% of the possible intergenic regions in each or-

ganism were used in the analysis, with no expectation that these are the same 60% between any two organisms, discovery of less than the complete set of snoRNAs is not unexpected.

More importantly, it should be remembered that *tricross* does not search specifically for snoRNAs. It was used here to search for shared sequences occurring in the *Pyrococcus* intergenic regions. snoRNA-gene-like sequences are simply the most obvious, and most easily verified feature in the set of shared sequences that *tricross* returned. To emphasize this point, we have highlighted several conserved sequences discovered by *tricross* in our online table at http://www.biosci.ohio-state.edu/~ray/tricross_psnos.html/.

These sequences, highly suggestive of snoRNA genes, are not included in early analyses of archaeal snoRNA content (Gaspin *et al.*, 2000), presumably because available assumptions regarding snoRNA structure and function precluded their candidacy. Of particular interest may be the sequence denoted psno17, which appears to target modification of tRNA-Gln (anticodon TTG) and tRNA-Gln (anticodon CTG), lending support to the findings of snoRNA targeted tRNA methylation in the Archaea. *Tricross* makes no assumptions regarding sequence structure or content, leaving to the researcher the task of determining the significance of any similar sequences found.

This paper has largely focussed on the application of *tricross* to the investigation of intergenic content, but as has been mentioned, the algorithm provides a generally applicable visualization method for a range of cross-genomic comparisons. To highlight this, we present Figure 5, which is the result of the application of *tricross* in two dimensions to the protein sequences of the three sequenced *Neisseria* genomes. In this analysis, *tricross* has been applied between pairs of genomes at the protein sequence level to detect proteins shared between them. This produces a picture very much like a classical dot-plot, displaying global genome ordering relationships, but in this case a plotted point indicates a shared pair of proteins, and the light or dark color of the point indicates whether a 2-way, or 1-way relationship was found between the sequences. In this case *tricross* produces a dot-plot visualization of the ordering relationship between the genomes, with dots linked to the aligned protein sequences. The results for the *N.meningitidis* Z2491 serogroup A strain compared with *N.gonorrhoeae* agree well with the published physical map for these organisms (Dempsey et al., 1995).

It is clear from these findings that *tricross* does produce the results for which it was designed. The application of dot-plot logic as a conserved-sequence finding, and noise-rejecting methodology quickly produces a collection of sequences worthy of further investigation. Application of the method in sequence-id space, rather than in nucleotide or residue sequence-coordinate space, allows the program to provide a convenient method for the examination of global genome organization, while uniquely tying each plotted 'dot' to named functional sequences within the genomes.

In the future, as more and more sequenced genomes become available, we expect *tricross* to be instrumental in assisting in multi-way comparisons of genome ordering and sequence conservation, as well as in providing insight into the fundamental question of what interesting sequences remain to be found between known genes.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Department of Energy, DE-FG02-91ER20041, to C.J.D. C.J.D is an associate of the Canadian Institute for Advanced Research.

This work was supported in part, by National Institutes of Health grant R01 AI45091 to R.S.M.

Credit is due in part to Dennis Maeder and his 3D CROSS program (Maeder, 1999) for the inspiration to attempt this mode of analysis.

We acknowledge the Gonococcal Genome Sequencing Project supported by USPHS/NIH grant #AI38399, and B.A.Roe, L.Song, S.P.Lin, X.Yuan, S.Clifton, Tom Ducey, Lisa Lewis and D.W.Dyer at the University of Oklahoma.

REFERENCES

- Armbruster,D.W. (1998) *Transfer RNA Maturation in the Archaea*, Dissertation, The Ohio State University.
- Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with Glimmer. *Nucleic Acids Res.*, **27**, 4636–4641.
- Dempsey,J.A., Wallace,A.B. and Cannon,J.G. (1995) The physical map of the chromosome of a serogroup A strain of *Neisseria meningitidis* shows complex rearrangements relative to the chromosomes of the two mapped strains of the closely related species *N.gonorrhoeae*. *J. Bacteriol.*, **177**, 6390–6400.
- Eddy,S.R. and Rivas,E. (2000) Secondary structure alone is not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
- Fleischmann,R.D. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
- Gaspin,C., Cavaille,J., Erauso,G. and Bachelier,J.-P. (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. *J. Mol. Biol.*, **297**, 895–906.
- Heilig,R. and Genoscope (1999) *Pyrococcus abyssi* genome sequence: insights into archaeal chromosome structure and evolution, GenBank genome description file.
- Kawarabayasi,Y. et al. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3. *DNA Res.*, **5**, 55–76.
- Kiss-Laszlo,Z., Henry,Y. and Kiss,T. (1998) Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J.*, **17**, 797–807.
- Lewis,L.A. et al. (2000) *Neisseria gonorrhoeae* strain FA1090 complete genomic sequence, <http://micro-gen.ouhsc.edu/>.
- Maeder,D.L. (1999) 3D Cross—*Pyrococcus* spp 3-way comparison, *Technical Report*, Frank Robb Lab, <http://comb5-156.umbi.umd.edu/poster/3Dcross/>.
- Omer,A.D., Lowe,T.M., Russell,A.G., Ehardt,H., Eddy,S.R. and Dennis,P.P. (2000) Homologs of small nucleolar RNAs in Archaea. *Science*, **288**, 517–522.
- Parkhill,J. et al. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, **404**, 502–506.
- Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
- Ray,W.C. and Daniels,C.J. (2001) The PACRAT system: an extensible WWW-based system for correlated sequence retrieval, storage and analysis. *Bioinformatics*, **17**, 100–104.
- Tettelin,H. et al. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**, 1809–1815.
- Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Utah (1999) *Pyrococcus furiosus* genome walk, <http://comb5-156.umbi.umd.edu/genemate/pfu-info.html>.
- VDC (1997) VRML 97: the virtual reality modeling language. ISO/IEC 14772-1:1997, *Technical Report*, VRML Consortium Inc., <http://www.vrml.org/technicalinfo/specifications/vrml97/index.htm>.