Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W659–W668***
doi:10.1093/nar/gkm334

DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities

F. Ferrè[1], Y. Ponty[2], W. A. Lorenz[2] and Peter Clote[2,*]

[1]Harvard Medical School, Children's Hospital, Hematology/Oncology Department, Boston, MA 02115 and [2]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA

Received January 31, 2007; Revised April 15, 2007; Accepted April 18, 2007

ABSTRACT

DIAL (dihedral alignment) is a web server that provides public access to a new dynamic programming algorithm for pairwise 3D structural alignment of RNA. DIAL achieves quadratic time by performing an alignment that accounts for (i) pseudo-dihedral and/or dihedral angle similarity, (ii) nucleotide sequence similarity and (iii) nucleotide base-pairing similarity.

DIAL provides access to three alignment algorithms: global (Needleman–Wunsch), local (Smith–Waterman) and semiglobal (modified to yield motif search). Suboptimal alignments are optionally returned, and also Boltzmann pair probabilities $Pr(a_i, b_j)$ for aligned positions a_i, b_j from the optimal alignment. If a non-zero suboptimal alignment score ratio is entered, then the semiglobal alignment algorithm may be used to detect structurally similar occurrences of a user-specified 3D motif. The query motif may be contiguous in the linear chain or fragmented in a number of non-contiguous regions.

The DIAL web server provides graphical output which allows the user to view, rotate and enlarge the 3D superposition for the optimal (and suboptimal) alignment of query to target. Although graphical output is available for all three algorithms, the semiglobal motif search may be of most interest in attempts to identify RNA motifs. DIAL is available at http://bioinformatics.bc.edu/clotelab/DIAL.

INTRODUCTION

During much of the 20th century the structural biology community has focused attention on the study of proteins, leading to a 'protein-centric' view of molecular and cellular biology, as manifest in various protein databases and tools: 'protein sequence' databases such as SwissProt (1), PIR (2), 'protein structure' databases such as the PDB (3), SCOP (4), CATH (5), tools such as PHD secondary structure prediction (6) and DALI structural alignment (7), etc.

In this century, RNA has emerged as an important focus of the structural biology community, as evidenced by the surprising and previously unsuspected roles played by RNA in genomic regulatory processes, such as post-transcriptional regulation with micro RNAs and small interfering RNAs (8), transcriptional and translational gene regulation by allosteric conformational changes in riboswitches (9,10), ribosomal frameshift induced by pseudoknots and slippery sequences (11) and chemical modification of specific nucleotides in the ribosome. Even the peptidyltransferase reaction in peptide bond formation is catalyzed by RNA (12,13).

Within this context, the current article describes the web server, DIAL (dihedral alignment), for pairwise structural alignment of RNA from input PDB files. Depending on the precise formulation of the problem, structural alignment of 3D protein/RNA backbone conformations is known to be *NP*-complete.[1] It follows that all current efficient algorithms either restrict the notion of 3D structural alignment or involve a heuristic. For protein structural alignment, DALI (7) and SSAP (Sequential Structure Alignment Program) (15) are

*To whom correspondence should be addressed. Tel: +1 617 552 1332; Fax: +1 617 552 2011; Email: clote@bc.edu

[1]In Lemma 3.3 of (14), Kolodny and Linial prove that ϵ-approximate optimal structural alignment is *NP*-complete, when the input consists of two distance matrices over an arbitrary metric space. Over 3D Euclidean space, Kolodny and Linial present an algorithm for ϵ-approximate optimal structural alignment of two proteins with run time $O(n^{10}/\epsilon^6)$.

perhaps the most widely used heuristic algorithms, where the latter has been used toward automatic classification in the CATH database (5).

With the current interest in RNA, ribonucleic acid sequence and structure databases have been established, and there is continual development of new algorithms and efficient software. For instance, Rfam (16) is an important sequence database of RNAs grouped by family (tRNA, SAM riboswitch, miRNA, etc.), while the NDB (Nucleic Acid Database) (17) is the primary repository for 3D RNA structures and the SCOR (Structural Classification of RNA) database is a derived data collection of RNA motifs (18).

There are far too many RNA alignment and motif-searching algorithms for us to properly survey the area in this article. To properly situate the contribution of DIAL, we give only the most necessary remarks. Most RNA structural alignment algorithms account for sequence and secondary structure similarity. In pioneering work, Sankoff (19) provided an important $O(n^6)$ algorithm to compute the optimal sequence and secondary structure alignment for two given RNA nucleotide sequences.[2] Both Foldalign (20) and Dynalign (21) are important practical implementations of reasonable restrictions of Sankoff's algorithm, using the Turner nearest neighbor energy model (22,23). Quite surprisingly, in the technical report (24) Blin *et al*. prove that optimal pairwise alignment is *NP*-complete, when the input consists of two RNA sequences along with their given secondary structures.[3] It follows that the precise stipulation of the input can effect the computational complexity of RNA structural alignment, a fact that explains in part the multitude of different algorithms for structural alignment.

In (27), Macke *et al*. describe the software RNAMotif developed for RNA motif search, allowing a flexible description of motif including any kind of base–base interaction. Liu *et al*. (28) present a quadratic time algorithm RSmatch for RNA secondary structure alignment and motif detection. Dalli *et al*. (29) describe the program STRAL, which performs a progressive alignment of non-coding RNA using base-pairing probability vectors in quadratic time. In an unusual approach, Sato and Sakakibara (30) apply conditional random fields to determine optimal RNA alignment.

Turning to RNA 3D structural alignment, in (31) Olson describes two virtual (or pseudo-) dihedral angles, later reintroduced by Duarte and Pyle (32). The pseudo-dihedral angles η respectively θ are determined by the four points $C4'(i-1), P(i), C4'(i), P(i+1)$ respectively $P(i), C4'(i), P(i+1), C4'(i+1)$, where $P(i)$ respectively $C4'(i)$ denotes the phosphorus atom respectively 4′-carbon atom of the ith RNA nucleotide. The program AMIGOS of Duarte and Pyle (32) computes RNA dihedral and pseudo-dihedral angles, used in the program PRIMOS of Duarte, Wadley and Pyle (33) to compute RNA 'worms', i.e. a sequence of η, θ angles for the entire RNA molecule. The method COMPADRES of Wadley and Pyle (34) uses PRIMOS to detect new RNA structural motifs, such as the π-turn, Ω-turn, α-loop, C2FA and hook turn (34), by the following procedure: (i) a non-redundant RNA structural data collection of 49 structures, 50 chains and 6697 nt is created; (ii) RNA worms are calculated for each of these structures, and the worms are concatenated into a single sequence; (iii) all maximal gapless matches of at least 5 nt of this sequence with itself are detected[4] and (iv) known 3D motifs are removed from the matches, and a frequency count is made of remaining matches, from which the high-frequency motifs are analyzed.

In (35), Hershkovitz *et al*. compute dihedral angles $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi$ and pseudorotational phase P for all nucleotides in the 3D structure of 23 S rRNA of *Haloarcula marismortui* with PDB ID 1S72:0. They identify similar contiguous sequences by 'torsion matching'; i.e. determining whether dihedral and pseudo-rotational angles differ by at most an angle-dependent threshold. Hershkovitz *et al*. then refine this analysis by binning the computed angles inorder to determine dihedral and pseudorotational angle preferences.

In (36,37), Dror, Nussinov and Wolfson describe a cubic time RNA tertiary structure alignment algorithm, ARTS, which proceeds by a seed match and greedy global extension to approximately compute the 'largest common point set' (LCP) between phosphorus atoms of two RNA molecules. Given PDB files for two RNA molecules A and B, the program ARTS first determines 3D coordinates $a_1, \ldots, a_n$ respectively $b_1, \ldots, b_m$ of all phosphorus atoms from A respectively B, then applies the software 3DNA of Lu and Olson to determine base pairs of each structure. Given RMSD error bound of ϵ, ARTS determines all seed matches of 'base quadrats'[5] $(i, i+1, j-1, j)$ and $(i', i'+1, j'-1, j')$ for which there is a rigid transformation (rotation and/or translation) T such that

$$\max\{||a_i - T(b_{i'})||, ||a_{i+1} - T(b_{i'+1})||, \\ ||a_{j-1} - T(b_{j'-1})||, ||a_j - T(b_{j'})||\} \leq \epsilon$$

Since there are $O(n)$ base pairs in an RNA molecule of length n, the computation of all seed matches is done in $O(n^2)$ time. Subsequently, a greedy extension of seed matches approximately computes the LCP $A' = \{a_{i_1}, \ldots, a_{i_k}\} \subseteq A$ and $B' = \{b_{i'_1}, \ldots, b_{i'_k}\} \subseteq B$ of phosphorus atoms between both RNA molecules, such that $||a_{i_x} - T(b_{i'_x})|| \leq \epsilon$ for all $1 \leq x \leq k$. The extension is done

[2]Sankoff provided a general $O(n^{3k})$ algorithm to determine the optimal multiple sequence/secondary structure alignment for k RNA nucleotide sequences of length n. To the best of our knowledge, there is no publicly available implementation of Sankoff's algorithm.

[3]In other words, the nested–nested edit-distance problem of Lin *et al*. (25) is *NP*-complete. See (26) for an $O(n^4)$ algorithm for a related RNA alignment problem.

[4]The worm $\langle(\eta_1, \theta_1), \ldots, (\eta_m, \theta_m)\rangle$ is defined to match the worm $\langle(\eta_1, \theta_1), \ldots, (\eta'_m, \theta'_m)\rangle$ if the Euclidean distance between (η_i, θ_i) and (η'_i, θ'_i) is at most 25° for each $1 \leq i \leq m$.

[5]A quadrat is a stack of size 2, i.e. positions i, $i+1$, $j-1$, j such that (i, j) and $(i+1, j-1)$ are base pairs.

to maximize a score $F(\ell, k) = w_1 \cdot k + w_2 \cdot \ell$, where there are ℓ base pairs and k nucleotides, for appropriate weights w_1, w_2. Note that ARTS does not necessarily respect the order of nucleotides in the linear chain, and that no account is taken for nucleotide identity; i.e. there is no nucleotide bonus for GNRA tetraloops.

In (38), Mokdad and Leontis describe the program Ribostral, which analyzes an RNA 3D alignment, and graphically presents base-pair isostericities. Sarver *et al.* (39) describe the algorithm FR3D used for RNA motif search by computationally intensive coordinate RMSD computations to determine optimal alignment between motif and target, by using a reduced atom representation of RNA nucleotides.

In this article, we introduce a quadratic time, dynamic programming algorithm, DIAL, able to find the optimal alignment with gaps (i.e. bulges) of two RNAs taking into account sequence, structure and base-pairing information extracted from the PDB file. DIAL provides a number of features not available in other RNA pairwise structural alignment algorithms. While PRIMOS and COMPADRES compute gapless alignments of pseudo-dihedral angles of contiguous segments, DIAL can perform global, local and semiglobal alignment in $O(n^2)$ time with affine gap penalty by taking into account nucleotide similarity,[6] dihedral and pseudo-dihedral angles as well as the base-pairing nature of nucleotides (0: unpaired, L: base paired with nucleotide to left, R: base paired with nucleotide to right). The program DIAL can perform alignments of 'fragmented' (i.e. non-contiguous or composite) motifs with targets, where the number of fragments is arbitrary. Since the computation of pseudo-dihedral angle η for the ith nucleotide requires atomic coordinates of both the $(i-1)$st and $(i+1)$st nucleotide, DIAL additionally extracts atomic coordinates from 1 nt preceding the start of the region specified and 1 nt following the end of the region specified. Inaccuracies (not checked by DIAL) will occur for the first and last nucleotide in the chain of a PDB file. The web server DIAL is an important extension of the program PRIMOS; indeed, PRIMOS alignments are obtained if DIAL parameters are specified to obtain a gapless alignment of pseudo-dihedral angles. This is done by entering negative gap initiation and gap extension parameters whose absolute value is prohibitively large, and setting to 0 all parameters for dihedral angles, nucleotide similarity, base-pairing nature (0,L,R). For user-specified 'suboptimal alignment score ratio' $0 \leq p \leq 1$, DIAL returns suboptimal alignments for which $\frac{S-S'}{S} \leq p$, where S denotes the optimal alignment score and S' denotes the suboptimal alignment score. Additionally, in quadratic time DIAL computes the partition function (41–43) for alignments, hence returns the Boltzmann pair probabilities $Pr(a_i, b_j)$ for aligned nucleotides a_i, b_j occurring in the optional alignment. Boltzmann pair probabilities can suggest the biological significance of portions of the optimal alignment, an idea validated for protein sequences by Vingron and Argos (44).

While ARTS is an excellent cubic time program for 'motif detection', yielding an approximation to the LCP set $A' = \{a_{i_1}, \ldots, a_{i_k}\} \subseteq A$ and $B' = \{b_{i'_1}, \ldots, b_{i'_k}\} \subseteq B$ of phosphorus atoms, it should be noted that ARTS does not necessarily preserve linear order within the alignment; i.e. it can happen that $i_j < i_\ell$ and $i'_j > i'_\ell$. Moreover, ARTS takes no account of nucleotide identity or similarity.

The graphical user interface of DIAL is particularly simple, in that PDB accession codes, chain IDs and starting and ending residue sequence numbers can be entered for both RNA molecules; optionally, PDB files can be uploaded. Allowing the user to fine-tune all parameters, DIAL is powerful, flexible and sufficiently accurate to allow the comparison of a large number of molecules for subsequent refinement by other methods.

MATERIALS AND METHODS

We computed RNA backbone dihedral angles by writing Python scripts based on the Biopython Structural Bioinformatics package (45). Six dihedral angles (α, β, γ, δ, ϵ, ζ) can be defined in the RNA backbone, and one dihedral angle (χ) describes the rotation between backbone and base. The values for all these angles are not independent, but there is a very high correlation between values of each pair of angles (35). Two additional virtual angles η and θ, first introduced by Olson (31) and later reintroduced by Duarte, Duarte, Wadley and Pyle (33) offer a reduced but sufficient conformational description of the RNA backbone (46). To determine base-pairing status of each nucleotide, we run RNAVIEW (47) on all the SCOR RNA chains.

Algorithm

In addition to quadratic time implementations of Needleman–Wunsch global alignment (48) and Smith Waterman local alignment (49) algorithms, the DIAL web server includes an implementation of 'semiglobal' alignment (50), opportunely modified to perform motif searching for contiguous or fragmented queries. All algorithms have been extended to account for the similarity of matched nucleotides, dihedral angles[7] and base-pairing attributes. To illustrate our modification of semiglobal alignment for fragmented motifs, suppose that the query consists of two non-contiguous fragments, $a_1, \ldots, a_m$ and $a'_1, \ldots, a'_{m'}$, and that the target consists of the contiguous sequence $b_1, \ldots, b_n$. In our semiglobal alignment, there is no penalty for gaps occurring to the left of a_1, between a_m and a'_1 and to the right of $a'_{m'}$, while gaps

[6]Default RNA nucleotide similarities are taken from BLAST (40); however, the user can modify nucleotide similarity, gap initiation and gap extension costs as well as other parameters.

[7]A dihedral or torsion angle is determined by four points a, b, c, d in Euclidean 3D space. By taking cross products, compute the normal vectors $\vec{u}, \vec{v}$ to the plane determined by a, b, c and b, c, d. The dihedral angle is defined to be the inverse cosine of the inner product $\langle \vec{u}, \vec{v} \rangle$ of $\vec{u}, \vec{v}$ normalized by their lengths.

occurring in $a_1, \ldots, a_m$ or in $a'_1, \ldots, a'_{m'}$ are penalized; i.e. alignment of the query is semiglobal. In contrast, all gaps in $b_1, \ldots, b_n$ are penalized, including those occurring to the left of b_1 and to the right of b_n. All algorithms run in quadratic time using affine gap penalties, following Gotoh's method (51).

Our scoring function evaluates the similarity between each nucleotide of the query and target, by accounting for (i) dihedral/pseudo-dihedral similarity, (ii) nucleotide sequence similarity and (iii) base-pairing similarity. Each one of these contributions is weighted by default parameters; these parameters can be modified by the user. In particular, if the user enters parameters x, y, z respectively for the dihedral, nucleotide and base-pairing parameters, then the weight of dihedral angle contribution is $x/(x+y+z)$, while that for nucleotide similarity is $y/(x+y+z)$, and that for base pairing is $z/(x+y+z)$. Similarly, one can modify the parameters for the seven dihedral angles and two pseudo-dihedral angles; i.e. if $x_1, \ldots, x_7$ respectively y, z denote the form values for the dihedral and pseudo-dihedral angles, then the first dihedral angle weight is $x_1/(x_1 + \cdots + x_7 + y + z)$, the weight for the first pseudo-dihedral angle η is $y/(x_1 + \cdots + x_7 + y + z)$, etc.

Given query $a_1, \ldots, a_n$ and target $b_1, \ldots, b_m$, the 'similarity' $\mathrm{sim}(a_i, b_j)$ of aligning a_i from the query RNA with b_j from the target RNA is given by the weighted sum

$$\mathrm{sim}(a_i, b_j) = w_1 \cdot \mathrm{Sequence}(a_i, b_j) + w_2 \cdot \mathrm{Backbone}(a_i, b_j) + w_3 \cdot \mathrm{BasePair}(a_i, b_j)$$

where, following BLAST default (40), the nucleotide sequence contribution $\mathrm{Sequence}(a_i, b_j)$ is 1 if nucleotides a_i, b_j are identical (match) and -3 otherwise (mismatch), and where

$$\mathrm{Backbone}(a_i, b_j) = \sum_{k \in A} \omega_k \cdot |k(a_i) - k(b_j)|.$$

Here A is the set of six backbone dihedral angles (α, β, γ, δ, ϵ, ζ), one dihedral angle (χ) describing the orientation of the base, and two pseudo-dihedral angles η respectively θ determined by the 4 points $C4'(i-1), P(i), C4'(i), P(i+1)$ respectively $P(i), C4'(i), P(i+1), C4'(i+1)$.[8] $\mathrm{BasePair}(a_i, b_j)$ is a penalty if the base-pairing attribute of a_i and b_j differ. Although we have focused discussion on the motif search application of DIAL, global and local 3D structural alignment is supported. Unless the parameters are set to be permissive, local alignment tends to report very small alignments of only a few nucleotides. Full details of the algorithm and extensions will be given in a forthcoming article.

Following Clote, Ferrè and Straubhaar (42,43), we additionally compute the Boltzmann pair probabilities within an optimal alignment (41) by computing a 'forward' Boltzmann partition function

$$FZ(i,j) = \sum_{\mathcal{A}} e^{\mathrm{sim}(\mathcal{A})/RT}$$

where $\mathcal{A}$ ranges over all possible alignments of $a_1, \ldots, a_i$ with $b_1, \ldots, b_j$, R is the universal gas constant and T is absolute temperature.[9] In the inductive case, the forward partition function $FZ(i,j)$ can be computed by

$$FZ(i-1,j-1) \cdot e^{\frac{\mathrm{sim}(a_i,b_j)}{RT}} + FZ(i,j-1) \cdot e^{\frac{\gamma}{RT}} + FZ(i-1,j) \cdot e^{\frac{\gamma}{RT}},$$

where for notational simplicity we have assumed a linear gap penalty γ.[10] In a similar fashion, the backward Boltzmann partition function BZ can be computed, where

$$BZ(i,j) = \sum_{\mathcal{A}} e^{\mathrm{sim}(\mathcal{A})/RT}$$

where $\mathcal{A}$ ranges over all possible alignments of $a_i, \ldots, a_n$ with $b_j, \ldots, b_m$. The Boltzmann probability $Pr[(a_i, b_j)]$ that a_i will be aligned with b_j is then

$$\frac{FZ(i-1,j-1) \cdot e^{\frac{\mathrm{sim}(a_i,b_j)}{RT}} \cdot BZ(i+1,j+1)}{FZ(n,m)}.$$

It should be stressed that due to the complexity of RNA 3D structural alignment, one cannot hope that a quadratic time algorithm such as DIAL be highly accurate. However, by using DIAL to compute potential target regions predicted to align well with the query, one can subsequently apply a very accurate, but computationally intensive RNA structural alignment algorithm, such as FR3D (39). We believe that this will be the primary application of DIAL.

Web server

The web server http://bioinformatics.bc.edu/clotelab/DIAL runs on a Linux cluster with 20 computational nodes, each with double processors of between 1300 and 3000 MHz and 2 GB RAM (6 Dell PowerEdge 1650, 2×1300 MHz Pentium III, 2 GB RAM; 11 Dell PowerEdge 1850, 2×2800 MHz Xeon EM64T, 2 GB RAM; 5 Dell PowerEdge 1850, 2×3000 MHz Xeon EM64T, 2 GB RAM).

The input form for DIAL is shown in Figure 1. The user must either upload or give the four character alphanumeric PDB accession code for both query and target RNA structures, and indicate the chain identifier for each (underscore if the PDB file contains no chain identifier). Optionally, the starting and ending residue sequence number for the query and/or target structure can be given. Default parameters for dihedral and pseudo-dihedral angle contributions to the alignment may be used or modified. The user can choose between the

[8]Given four points, *a, b, c, d,* the first three and last three determine two planes. The dihedral angle between the planes is computed by taking the inverse cosine of the inner product of the normal to each plane.

[9]In alignment, temperature is a non-physical parameter; however, as in (42), by taking several temperatures one sees the overall significance of portions of the alignment (44).

[10]DIAL uses a general affine gap penalty, following Gotoh s algorithm (51).

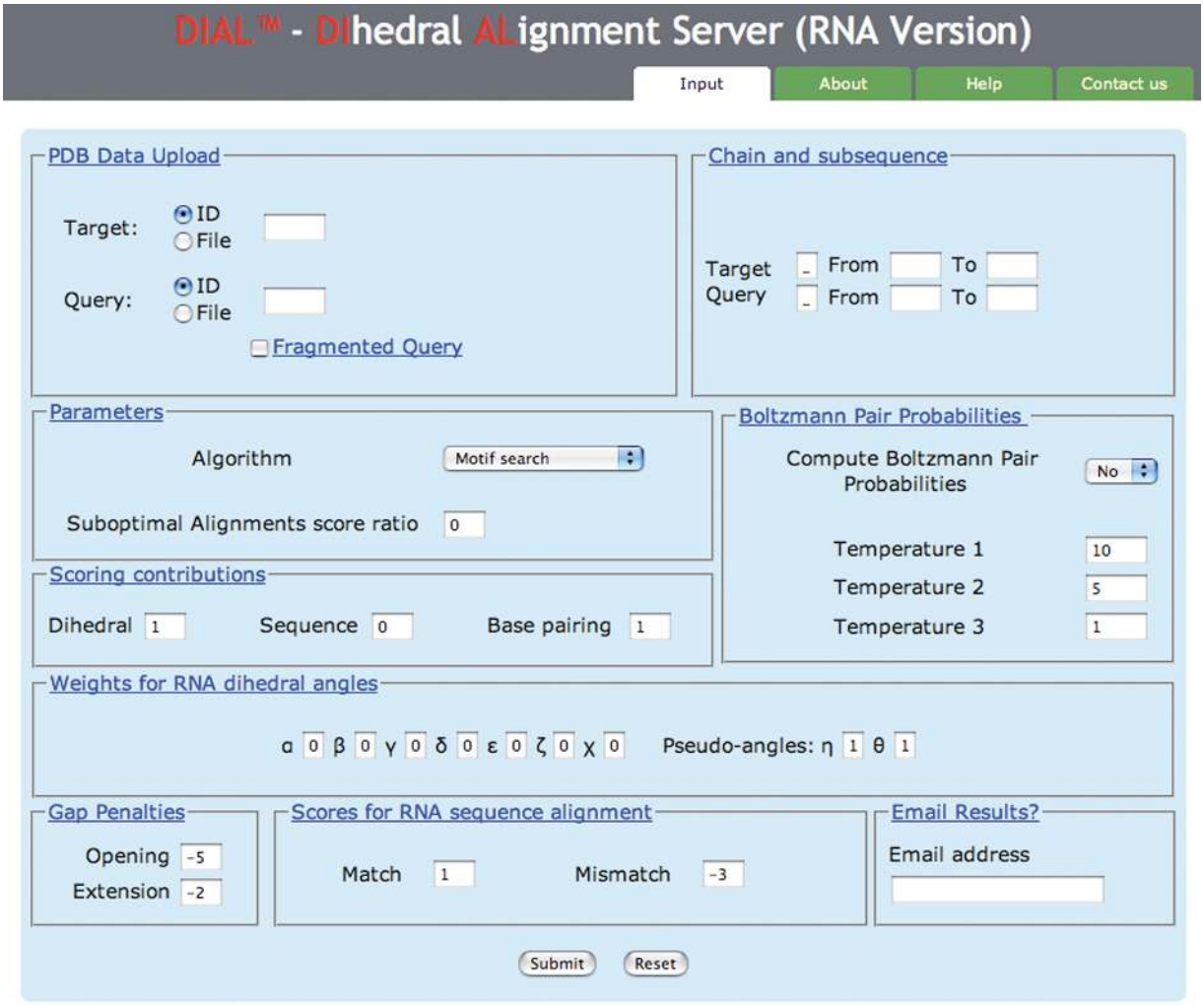

Figure 1. DIAL input form. The user must either upload query and target PDB files, or give the four character PDB code, and additionally indicate the chain identifier (underscore indicates a blank chain identifier in the PDB file). By modifying the parameter α, the user can appropriately weigh the sequence versus dihedral angle contribution to the alignment score. By default, the alignment takes into account only the two pseudo-dihedral angles η, θ and the base-pairing similarity.

semiglobal motif finding algorithm (default), or global or local alignment. Three temperatures may be chosen for the Boltzmann pair probability computation to determine highly significant portions of the alignment. Figure 2 displays the output of DIAL, when executing semiglobal alignment of query 1J5A (chain A, nucleotides 2530–2536) with target 1HR2 (chain A). Hot links are provided for the alignment, dihedral and pseudo-dihedral angles (and sugar pucker), Boltzmann probabilities and superposition; alignment and a zoomed close-up of the superposition are depicted in Figure 3.

RESULTS

To illustrate the difference in alignment accuracy of DIAL and ARTS, we applied the motif search algorithm to two transfer RNA structures. The query structure was 1ASZ:R from residue sequence number 620 to 660 and the target structure was 4TRN. While DIAL correctly aligned this 41 nt portion of aspartyl-tRNA 1ASZ with the corresponding portion of 4TRN, the alignment produced by ARTS is incorrect; see Figure 4. For certain examples, this comportment of ARTS is not surprising, since it was designed to compute the largest collection of phosphorus atoms which are ϵ-close to each other.

To assess the accuracy of the DIAL web server, we computed receiver operating characteristic (ROC) curves (52), which depict the trade-off between sensitivity (true positive rate) and specificity (1 minus false positive rate). For this assessment, we used the SCOR database (18).

SCOR XML dumps were parsed in order to locally reconstruct the SCOR database. Our starting structure data set included all RNA motifs in the SCOR database; i.e. 440 families and altogether 9850 motifs. Of 440 SCOR families, 82 had both fragmented and non-fragmented members while 62 had only fragmented members. Note that even if the number of SCOR families having fragmented members is relatively small, they are often the most populated families. There were 5110 members in families having 2 fragments, 3 members in families having 3 fragments, 1 member in a family of 4 fragments

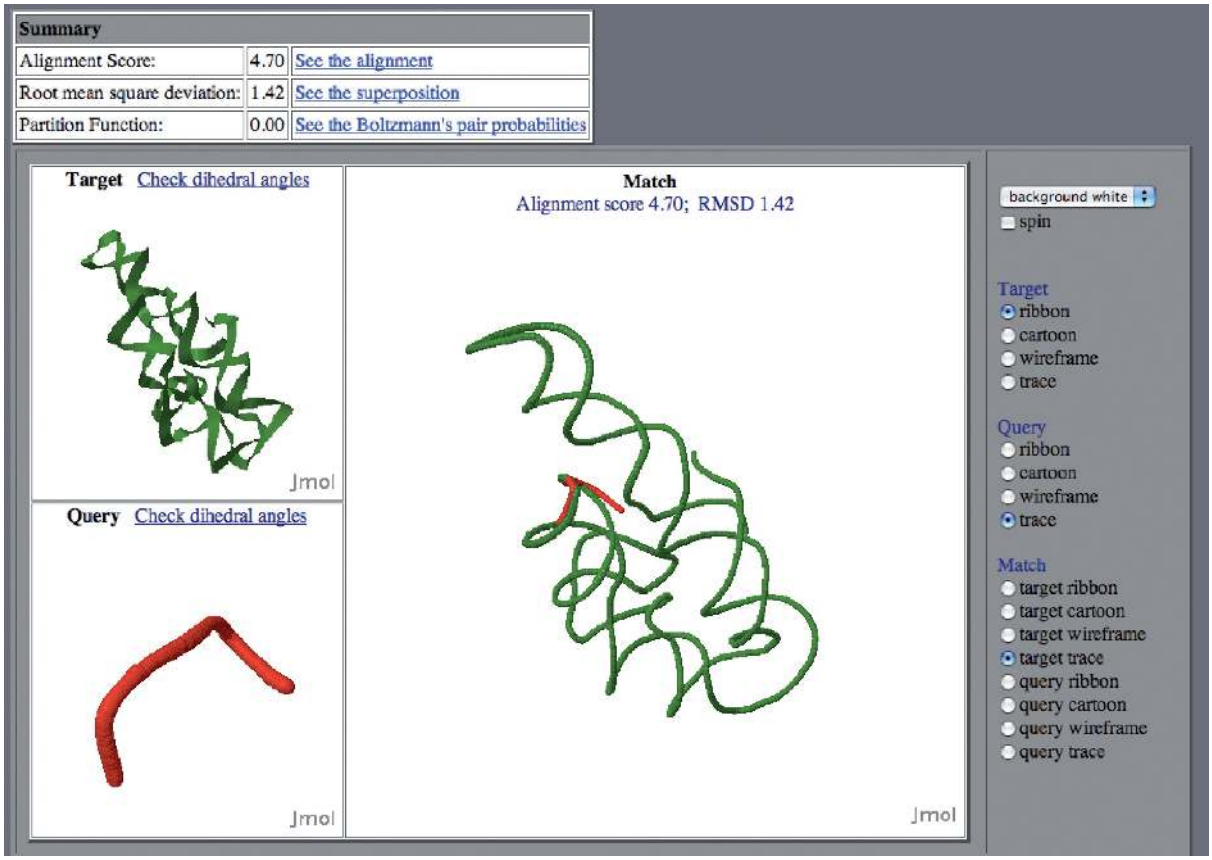

Figure 2. DIAL screen output, when applying the motif detection (semiglobal alignment) algorithm of query 1J5A (chain A, nucleotides 2530–2536) with target 1HR2 (chain A). The target (respectively query) conformation is depicted in the upper (respectively lower) left corner, along with hot links to the computed dihedral angles. The superposition of optimal query to target alignment is depicted on the right. The images are produced by using a JMOL applet, hence allow the user to rotate, zoom in, zoom out and choose a variety of molecule representations. To the right of this output (not shown) is a pull-down tab for suboptimal alignments, provided the user entered a non-zero parameter for suboptimal alignment score ratio.

and 1 member in a family of 6 fragments. We filtered the SCOR collection to eliminate the following motifs: (i) shorter than 3 nt; (ii) composite motifs where fragments belong to different chains and (iii) no range is specified. (i.e. starting and ending position in the chain.) After this selection, we accepted only SCOR families having more than one remaining motif. This step produced 136 families and altogether 5619 motifs. Of these 136 families, 89 contained only local (contiguous, non-fragmented) motifs, 41 contained only composite (fragmented) motifs, and 6 contained both local and composite examples. Of a total of 5619 motifs, 2836 are composite, all formed by two fragments.

Since SCOR includes RNA structures which may be identical or very similar but have different PDB accession codes, for each SCOR family we produced a sequence non-redundant subcollection using Algorithm 2 described in (53).[11] In this process, we additionally discarded structures shorter than 5 nt and having poorer resolution than 3.5 Å. Our final, filtered, non-redundant data set extracted from SCOR database thus consisted of 78 families and altogether 359 motifs. The reason there were so few remaining motifs is due to the fact that the SCOR database has many identical or very similar motifs occurring in different RNA molecules.

Figure 5 and 6 present ROC curves respectively for contiguous and fragmented queries. These are computed as follows. For each pair (S_1, S_2) of structures in the non-redundant data collection obtained from SCOR as indicated above, we computed the DIAL similarity

$$\mathrm{sim}(S_1, S_2) = w \cdot \mathrm{seqSim}(S_1, S_2) + (1 - w) \cdot \mathrm{strSim}(S_1, S_2) + \mathrm{bpSim}(S_1, S_2)$$

where seqSim represents nucleotide sequence similarity, strSim represents pseudo-dihedral η, θ angle similarity and

[11]Algorithm 2 constructs a sequence non-redundant data set as follows. Given list L of sequences, determine BLAST similarity of first sequence to all others, removing from L all homologous sequences (with E-value above a given threshold). Take the second sequence from the filtered list L, determine BLAST similarity with all successive sequences from L, removing those which are homologous, etc. In this fashion a set of sequences is obtained, guaranteed not to be pairwise homologous. In our implementation, we used default BLAST values for nucleotide match, mismatch and gap, set the threshold to be the E-value 0.001.

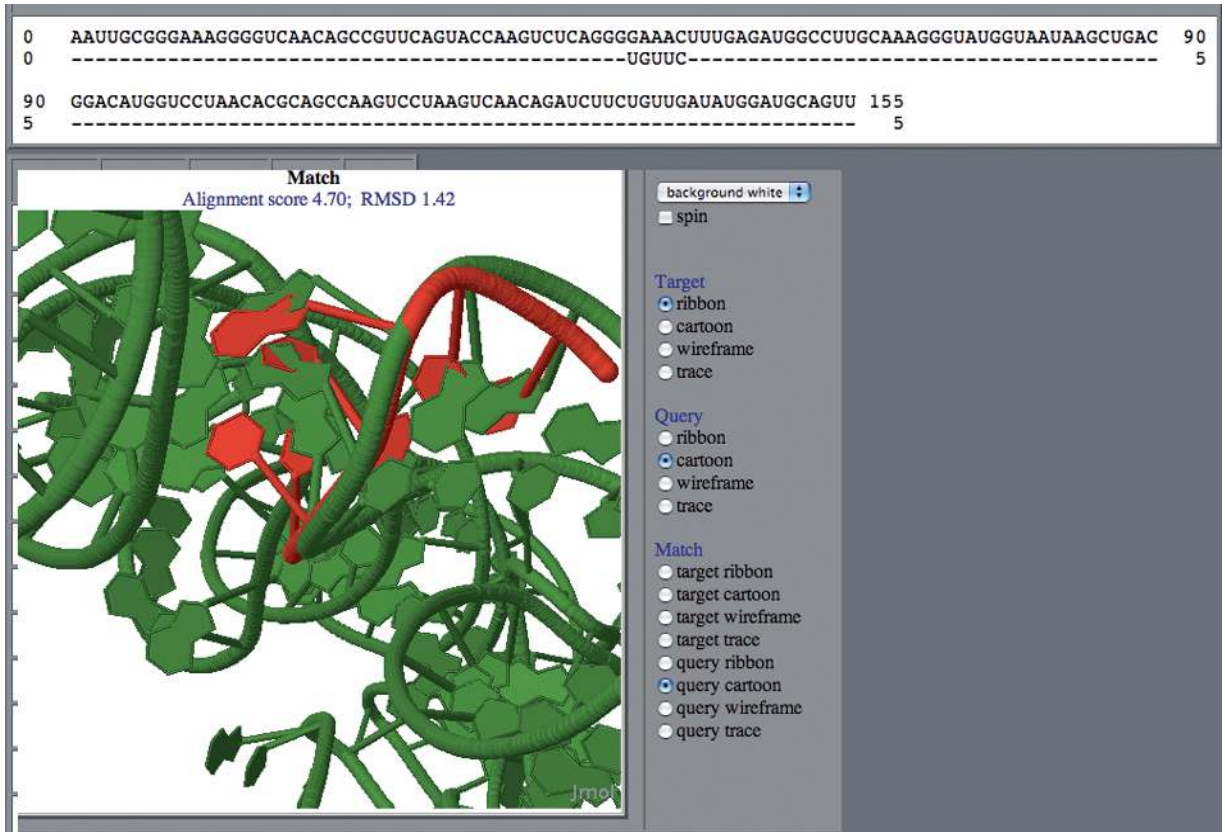

Figure 3. (Top) Optimal alignment produced in this case by the semiglobal alignment of query to target, when applying the motif detection (semiglobal alignment) algorithm of query 1J5A (chain A, nucleotides 2530–2536) with target 1HR2 (chain A). Output includes a computation of the Boltzmann pair probabilities (not shown). (Bottom) An enlarged superposition of query to target; user can rotate and zoom in/out of image, and choose various representations of both query and target.

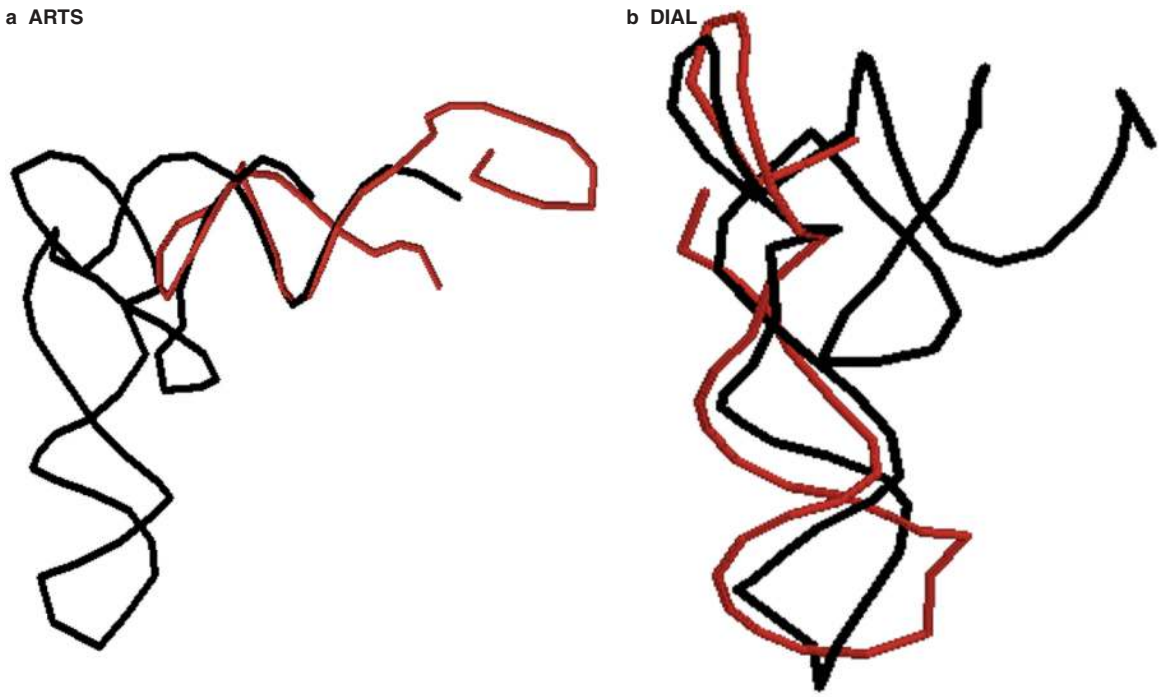

Figure 4. Alignment of contiguous fragment of aspartyl-tRNA 1ASZ:R starting from residue sequence number 620 and ending with 660 with the tRNA 4TRN. Left panel displays the first alignment produced by ARTS; right panel displays output of DIAL using the motif alignment algorithm with default parameters.

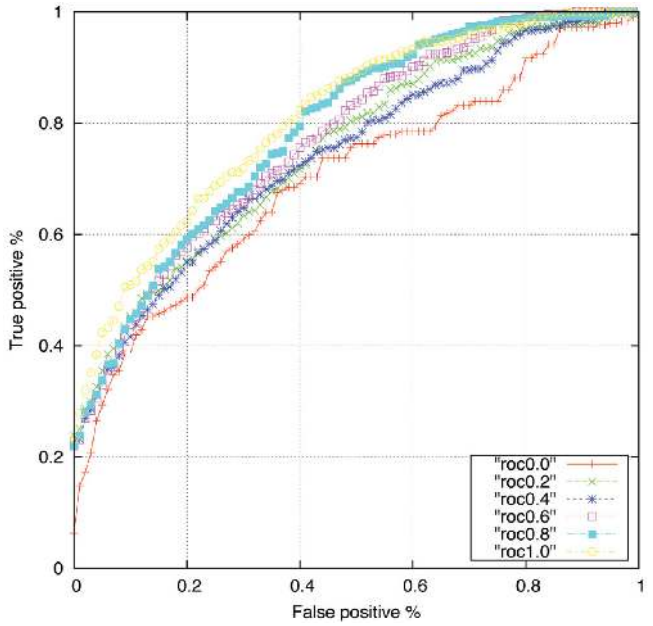

Figure 5. Average ROC curves when using the semiglobal DIAL algorithm to align query motifs from the SCOR database with targets from the SCOR database. The *x*-axis represents false positive rate (1 minus specificity), while the *y*-axis represents true positive rate (sensitivity). Overlaid curves represent different weighting of dihedral angle versus sequence contributions with weights $w = 0, 0.2, \ldots, 0.8, 1.0$. (See Table 1 or text for fuller description of parameters used.) This figure depicts ROC curves for contiguous queries, consisting of an uninterrupted linear sequence of nucleotides.

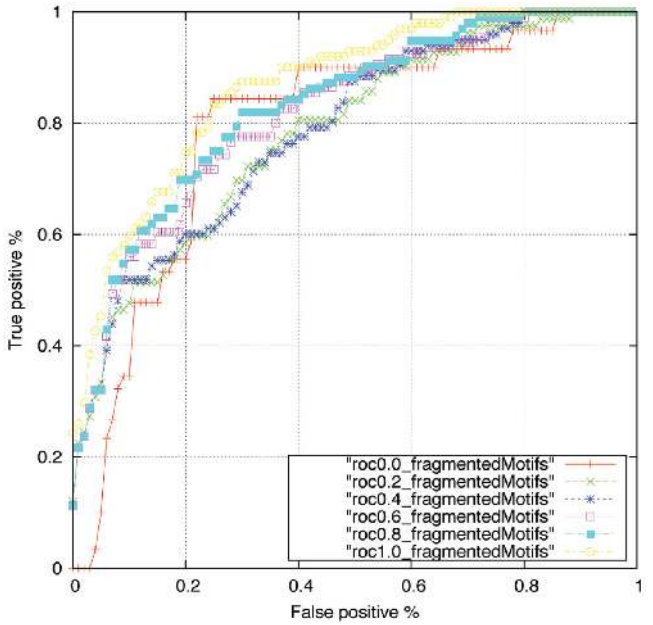

Figure 6. Average ROC curves when using the semiglobal DIAL algorithm to align query motifs from the SCOR database with targets from the SCOR database. The *x*-axis represents false positive rate (1 minus specificity), while the *y*-axis represents true positive rate (sensitivity). Overlaid curves represent different weighting of dihedral angle versus sequence contributions with weights $w = 0, 0.2, \ldots, 0.8, 1.0$. (See Table 1 or text for fuller description of parameters used.) This figure depicts ROC curves for fragmented queries, representing 3D motifs consisting of two or more interrupted linear sequences of nucleotides. In the SCOR database, most fragmented queries consist of two contiguous linear sequences. (See text for fragment breakdown for the SCOR databse.)

bpSim represents base-pairing similarity.[12] Computations were performed for weights w from $0, 0.2, 0.4, \ldots, 0.8, 1.0$. Positives (respectively negatives) were considered pairs (S_1, S_2) from the same (respectively different) SCOR class. This allowed the computation of ROC curves displayed in Figure 5. For the most part, pseudo-dihedral angle similarity is much more important for proper SCOR classification than nucleotide sequence similarity.

These data gave rise to the ROC curves shown in Figure 6, which displays overlaid curves with different weights w for the sequence versus structural alignment, $w = 0, 0.2, 0.4, \ldots, 0.8, 1.0$. Table 1 presents the area under ROC curves, denoted by AUC, for both non-fragmented and fragmented motifs, using the data from the previously discussed ROC curves.

Table 1. Area under ROC curve (AUC) for ROC curves displayed in Figures 5 and 6. ROC curves were created for a non-redundant data set extracted from the SCOR database—see Section Data set in reference (20). AUC is computed for different values of weight parameter w for both non-fragmented and fragmented queries, for $w = 0, 0.2, \ldots, 0.8, 1.0$. This corresponds to setting parameters on DIAL web form as follows: 'dihedral' $= w$, 'sequence' $= (1-w)$, 'base-pairing' $= 1$. With these settings, DIAL alignments give same weight to sequence/structural similarity and base-pairing similarity. By varying weight w, we obtain a trade-off between sequence and dihedral angle similarity. With these settings, DIAL appears to perform slightly better on fragmented motifs

w	Non-fragmented AUC	Fragmented AUC
0.0	0.69	0.78
0.2	0.74	0.77
0.4	0.73	0.78
0.6	0.76	0.81
0.8	0.78	0.82
1.0	0.80	0.86

DISCUSSION

In this article, we have described the DIAL web server, which provides access to global, local and semiglobal alignment of RNA structures, presented as PDB files. We believe the semiglobal alignment to be of particular interest as a preprocessing step for RNA motif detection.

The DIAL web server performs a quadratic time, dynamic programming alignment, taking into account similarity of nucleotide identity, (pseudo-) dihedral angles and base pairing in the secondary structure. The algorithm is fully customizable by allowing the user to stipulate different weights for angles and base pairs.

[12]set parameters on web form as follows: 'dihedral' $= w$, 'sequence' $= (1-w)$, 'base pairing' $= 1$. Dihedral angle parameters α through χ are set to 0, while parameters η and θ are set to 1.

Unlike the PRIMOS algorithm of Duarte *et al.* (33), which considers a gapless alignment of pseudo-dihedral angles for contiguous sequences, DIAL can handle fragmented queries and alignments with bulging nucleotides by means of gap insertion. DIAL alignment accounts for base-pairing similarity, known to be of primary importance in the manual curation of the SCOR database. Additionally, DIAL computes Boltzmann pair probabilities in the alignment, and can return suboptimal query-target alignments.

ACKNOWLEDGEMENTS

Research of P.C., W.A.L. and Y.P. was supported by National Science Foundation grant DBI-0543506. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We wish to thank Steve Holbrook for pointing out the importance of properly accounting for base pairing in an earlier version of DIAL, Jason Persampieri for technical assistance and both N. Leontis and C. Zirbel for generously providing us with a preprint of the article (39). Funding to pay the Open Access publication charges for this article was provided by National Science Foundation grant DBI-0543506.

Conflict of interest statement. None declared.

REFERENCES

1. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
2. Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z.Z., Ledley,R.S., Lewis,K.C., Mewes,H.W. *et al.* (2002) The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.
3. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr.*, **58**, 899–907.
4. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
5. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo, C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
6. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
7. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
8. Lim,L., Glasner,M., Yekta,S., Burge,C. and Bartel, D. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
9. Winkler,W.C., Cohen-Chalamish,S. and Breaker,R.R. (2002) An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci. USA*, **99**, 15908–15913.
10. Penchovsky,R. and Breaker,R. (2005) Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nat. Biotechnol.*, **23**, 1424–1431.
11. Bekaert,M., Bidou,L., Denise,A., Duchateau-Nguyen,G., Forest,J., Froidevaux,C., Hatin,I., Rousset,J. and Termier, M. (2003) Towards a computational model for −1 eukaryotic frameshifting sites. *Bioinformatics*, **19**, 327–335.
12. Weinger,J., Parnell,K., Dorner,S., Green,R. and Strobel,S. (2004) Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nat. Struct. Mol. Biol.*, **11**, 1101–1106.
13. Nissen,P., Hansen,J., Ban,N., Moore,P. and Steitz,T. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–923.
14. Kolodny,R. and Linial,N. (2004) Approximate protein structural alignment in polynomial time. *Proc. Natl Acad. Sci. USA*, **101**, 12201–12206.
15. Taylor,W.R., Flores,T.P. and Orengo, C.A. (1994) Multiple protein structure alignment. *Protein Sci.*, **3**, 1858–1870.
16. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
17. Berman.,HM., Westbrook J., Feng Z., Iype L., Schneider B., Zardecki C. (2003) The nucleic acid database. *Methods Biochem Anal.*, **44**, 199–216.
18. Klosterman,P., Tamura,M., Holbrook,S. and Brenner, S. (2002) SCOR: a structural classification of rna database. *Nucleic Acids Res.*, **30**, 392–394.
19. Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
20. Havgaard,J.H., Lyngso,R.B. and Gorodkin,J. (2005) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.*, **33**, W650–W653.
21. Mathews, D. and Turner, D. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
22. Matthews,D., Sabina,J., Zuker,M. and Turner,D. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
23. Xia,T., SantaLucia,J., Burkard,M., Kierzek,R., Schroeder,S., Jiao,X., Cox,C. and Turner, D. (1999) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.
24. Blin,G., Fertin,G., Rusu,I. and Sinoquet,C. (2003) RNA sequences and the EDIT(NESTED,NESTED) problem. *Technical Report 03.07 Research Report*. Submitted for publication.
25. Jiang,T., Lin,G., Ma,B. and Zhang.K. (2002) A General Edit Distance between Two RNA Structures. *Journal of Computational Biology* 9(2): 371–388.
26. Herrbach,C., Denise,A., Dulucq,S. and Touzet,H. (2006) Alignment of RNA secondary structures using a full set of operations Technical Report 1451 Laboratoire de recherche en informatique (LRI). *Research Report*. Submitted for publication.
27. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
28. Liu,J., Wang,J.T., Hu,J. and Tian,B. (2005) A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC. Bioinformatics*, **6**, 89.
29. Dalli,D., Wilm,A., Mainz,I. and Steger,G. (2006) STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, **22**, 1593–1599.
30. Sato, K. and Sakakibara,Y. (2005) RNA secondary structural alignment with conditional random fields. *Bioinformatics*, **21**, ii237–ii242.
31. Olson,W.K. (1975) Configurational statistics of polynucleotide chains. A single virtual bond treatment. *Macromolecules*, **8**, 272–275.
32. Duarte,C. and Pyle,A. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. *J. Mol. Biol.*, **284**, 1465–1478.
33. Duarte,C.M., Wadley,L.M. and Pyle,A.M. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, **31**, 4755–4761.
34. Wadley,L. and Pyle,A. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated

approach to structural discovery. *Nucleic Acids Res.*, **32**, 6650–6659.
35. Hershkovitz,E., Tannenbaum,E., Shelley,B., Sheth,A., Tannenbaum,A. and Williams,L. (2003) Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Res.*, **31**, 6249–6257.
36. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21**, ii47–ii53.
37. Dror,O., Nussinov,R. and Wolfson,H.J. (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res.*, **34**, W412–W415.
38. Mokdad,A. and Leontis,N.B. (2006) Ribostral: an RNA 3D alignment analyzer and viewer based on basepair isostericities. *Bioinformatics*, **22**, 2168–2170.
39. Sarver,M., Zirbel,C., Stombaugh,J., Mokdad,A. and Leontis,N. (2006) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures, *J. Math Biol.*
40. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
41. Mückstein,U., Hofacker,I. and Stadler,P. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18**, S153–S160.
42. Clote,P. and Straubhaar,J. (2006) Symmetric time warping, Boltzmann pair probabilities and functional genomics. *J. Math. Biol.*, **53**, 135–161.
43. Ferre,F. and Clote,P. (2006) BTW: a web server for Boltzmann time warping of gene expression time series. *Nucleic Acids Res.*, **34**, W482–W485.
44. Vingron,M. and Argos,P. (1990) Determination of reliable regions in protein sequence alignments. *Protein Eng.*, **3**, 565–569.
45. Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
46. Wadley,L.M. and Pyle,A.M. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res.*, **32**, 6650–6659.
47. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
48. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
49. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
50. Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.
51. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
52. Gribskov,M. and Robinson,N. (1996) The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–34.
53. Hobohm,U., Scharf,M., Schneider,R. and C.Sander (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Proteins Sci.*, **1**, 409–417.