BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 17 2011, pages 2443–2445
doi:10.1093/bioinformatics/btr417

Structural bioinformatics

Advance Access publication July 14, 2011

NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments

J. Y. Semegni[1,*], M. Wamalwa[2], R. Gaujoux[1], G. W. Harkins[2], A. Gray[1] and D. P. Martin[1]

[1]Computational Biology Group, Department of Clinical Laboratory Sciences, IIDMM, University of Cape Town, 7925 Anzio Road, Observatory and [2]South Africa National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville 7535, Cape Town, South Africa

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Many natural nucleic acid sequences have evolutionarily conserved secondary structures with diverse biological functions. A reliable computational tool for identifying such structures would be very useful in guiding experimental analyses of their biological functions. NASP (Nucleic Acid Structure Predictor) is a program that takes into account thermodynamic stability, Boltzmann base pair probabilities, alignment uncertainty, covarying sites and evolutionary conservation to identify biologically relevant secondary structures within multiple sequence alignments. Unique to NASP is the consideration of all this information together with a recursive permutation-based approach to progressively identify and list the most conserved probable secondary structures that are likely to have the greatest biological relevance. By focusing on identifying only evolutionarily conserved structures, NASP forgoes the prediction of complete nucleotide folds but outperforms various other secondary structure prediction methods in its ability to selectively identify actual base pairings.

Availability: Downloable and web-based versions of NASP are freely available at http://web.cbio.uct.ac.za/~yves/nasp_portal.php

Contact: yves@cbio.uct.ac.za

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 25, 2011; revised on June 14, 2011; accepted on July 9, 2011

Besides a capacity to store information within the sequences of their component nucleotides, single-stranded nucleic acids may also store information within their secondary structures. Under physiological conditions many single-stranded RNA or DNA molecules longer than approximately 20 nucleotides form meta-stable secondary structures, which can have important roles in genome replication and gene expression. Although a number of computational methods exist for predicting nucleic acid secondary structures from either single sequences or alignments (Bernhart *et al*., 2008; Hamada *et al.,* 2009; Knudsen and Hein, 2003; Markham and Zuker, 2008), even the best of these incorrectly infer a high proportion of base pairings. Also, only a few methods provide any measures of statistical support either for their folding predictions, or for the overall presence of secondary structure (Babak *et al*., 2007; Simmonds *et al*., 2004). From the perspective of experimental biologists seeking to test the functional relevance of secondary structures, it would be very useful to have a computational tool that, with the lowest possible false positive rate, will identify sites that pair within evolutionarily conserved secondary structures.

NASP is an attempt to improve the selectivity with which individual secondary structures can be identified. It uses base pairing probabilities provided by the UNAfold nucleic acid folding program hybrid-ss (Markham and Zuker, 2008) that applies a combined partition function calculation, stochastic sampling and dynamic programming approach to compute base pairing probabilities and minimum free energy (MFE) estimates from single-stranded nucleotide sequences. The rationale behind NASP is simple: we assume that randomly shuffling nucleotides within sequences that have evolved to form stable secondary structures should influence their overall base pairing potential such that the shuffled sequences should yield higher MFE estimates than the real sequences from which they were produced. By comparing MFE estimates made with real sequences to those made with randomized versions of these sequences, NASP tests whether there is evidence that the real sequences have greater structure forming capability than can be accounted for by chance.

For each sequence, k, in an input alignment, hybrid-ss estimates the over-all Gibbs free energy of an optimally folded nucleotide sequence and yields a list of Boltzmann probabilities $P_k(i, j)$ of individual potential base pairings. NASP then computes a consensus base pairing matrix, M, whose entries satisfy

$$M_{ij} = 2^{Cij}\ \delta_{ij} \sum_{k}^{N} w_k\ \log(P_k(i,j)/P_T)$$

where: N is the number of sequences; $P_k(i, j)$'s are chosen to be above a user specified threshold probability P_T; w_k is the mean pair-wise Hamming distance of sequence k from all others in the alignment, such that w_k weighs the contribution to M_{ij} of more divergent sequences more heavily than those of less divergent sequences; δ_{ij} is 0 if the gap frequency at either position i or j is ≥ 0.75 and 1 otherwise; $C_{ij,}$ is the mutual information of columns i and j and accounts for compensatory mutations. It is given by

$$C_{ij} = -\sum_{(a,b)} f_{ij}(a,b)\log_2(f_{ij}(a, b)/f_i(a)f_j(b))$$

*To whom correspondence should be addressed.

Table 1. NASP compared with other RNA folding programs using sequences with known folds (best scores in bold)

Dataset	Length	#seq[a]	MPI[b]	NASP						RNAalifold			CentroidAliFold			Pfold			EvoFold		
				Pre[c]			Post[d]														
				Sel	MCC	FP[e]	Sel	MCC	FP	Sel	MCC	FP	Sel	MCC	FP	Sel	MCC	FP	Sel	MCC	FP
Corona_pk3	63	10	H	**1**	**1**	**0**	**1**	0.84	**0**	**1**	0.89	0	**1**	**1**	**0**	0.63	0.55	3	0.64	0.67	2
Corona_pk3	63	10	H	**1**	**1**	**0**	**1**	0.84	**0**	**1**	0.89	0	**1**	**1**	**0**	0.63	0.55	3	0.64	0.67	2
Hammerhead_1	45	1		0.71	0.77	5	**1**	0.70	**0**	0.83	**0.83**	3	0.83	**0.83**	3	0.82	0.78	3	ND	ND	ND
Purine	102	4	H	**1**	**0.95**	**0**	**1**	0.80	**0**	**1**	**0.95**	**0**	**1**	**0.95**	**0**	**1**	0.77	**0**	0.66	**0.95**	**1**
HIV-1 NL4-3	9173	1		0.30	0.32	1556	0.46	0.29	**399**	0.31	0.35	1684	**0.49**	**0.44**	421	ND[f]	ND	ND	ND	ND	ND
S2m	43	38	M	0.62	0.63	5	**1**	0.64	**0**	**1**	**1**	**0**	0.82	0.72	2	0.82	0.78	2	0.25	0.27	12
SSU_rRNA	1542	11	M	0.72	**0.72**	128	**0.84**	0.68	**51**	0.67	0.63	135	0.66	0.59	128	ND	ND	ND	ND	ND	ND

[a]The number of sequences in the alignment.
[b]The mean pair-wise sequence identity. M stands for medium similarity (MPI between 60% and 90%) and H for high similarity (MPI between 80% and 99%).
[c]NASP performances before enrichment and permutation testing. The numbers in bold represent the best scores.
[d] NASP performance after permutation testing by mononucleotide shuffling.
[e]The number of falsely predicted base-pairs.
[f] Data is unavailable due to restrictions on input sequence lengths and number.

where; $f_i(a)$ is the frequency of nucleotide a (= A, C, G or T) at alignment position i; $f_{ij}(a, b)$ is the frequency of finding nucleotide a at position i and b at position j in the alignment. $C_{ij} = 0$ if columns i and j have evolved independently and is > 0 otherwise. Thus during the calculation of M_{ij} the factor 2^{Cij} is used to weigh co-evolving base pairs more heavily than other sites.

NASP aims to find a set of base pairs $i.j$ that maximizes $\sum M_{ij}$. To avoid alignment gaps obscuring signals of conserved structural motifs within M we allow relaxation on the requirement that homologous nucleotides must fall within the same alignment column. Specifically, we consider nucleotides within different sequences to be potentially homologous if they are separated by no more than d bases within the alignment (where d is a user-specified non-negative integer, usually between 0 and 10). Whereas if d = 0 only nucleotides within the same alignment column will be considered homologues, if d = 10 all nucleotides falling within 10 alignment columns up- and down-stream of a nucleotide column will be considered its potential homologues. This translates into

$$M_{ij} = \sum_{|r-i| \le d} \quad \sum_{|s-j| \le d} M_{rs}$$

NASP scans M through the anti-diagonal and recursively identifies groups of potentially base paired nucleotides displaying the highest degree of evolutionary conservation (i.e. contiguous non-zero entries in M that have the highest sum). At each step:

(1) The coordinates of nucleotides within the bounds of what appear to be the largest and most evolutionarily conserved structure represented in M are added to a list of potentially paired sites (Supplementary Figure S1).

(2) All alignment columns that are not included in this list are randomly shuffled 100 or more times with the MFEs of each sequence in each shuffled alignment being compared with those of sequences in the original alignment.

(3) The existence of additional unaccounted for structural motifs in the sequences is inferred when the MFE estimates of all sequences in the unshuffled alignment are smaller than those of at least 95% of the corresponding sequences in the shuffled alignments. The probability that there remain no unaccounted for paired nucleotides within the alignment fraction excluded from the potentially paired site list is estimated as the fraction of shuffled sequences with MFE estimates lower than those of their unshuffled counterparts.

(4) When the MFEs of the unshuffled sequences are less than those of 95% of their shuffled counterparts, the recursion continues from (1) with sites in the next most evolutionarily conserved structure being added to the paired site list.

The time complexity of NASP, which should be substantially reduced given that MFEs are computed in parallel, is $O(NPL^3)$ and the space complexity is $O(L^2)$ where L is the length, N the number of sequences in the alignment and P the number of permutation tests. Currently our web version of NASP accepts datasets of, for example 30×3 Kb sequences or 10×10 Kb sequences and analyses these within 144 h (results from comparison of computational times are shown in Supplementary Table S1).

Given a sequence alignment containing evidence of evolutionarily conserved secondary structures as input, NASP outputs (i) the coordinates of potentially conserved stems and *P*-values indicating statistical support for additional unaccounted for secondary structures remaining in the sequences following each recursion, (ii) the consensus structure in both the Vienna bracket-dot and a concatenation file formats and (iii) the consensus base-pairing matrix, M, in both text and graphical formats.

Tests using known reference RNA structures (Table 1 for examples and Supplementary Table S1 for the complete set) indicate that the overall selectivity (the proportion of inferred base pairs that are actually in the reference structures) of NASP is considerably better than that of RNAalifold (Bernhart *et al*., 2008), Pfold (Knudsen and Hein, 2003), CentroidAliFold (Hamada *et al*., 2009) and EvoFold (Pedersen *et al.,* 2006). The cost of NASP's low false positive rate is, however, a decreased true positive rate such that its overall accuracy (measured here using the Mathews Correlation Coefficient, MCC of Gardner and Giegerich, 2004) is slightly lower

than that of RNAalifold (which was overall the most accurate of the programs we tested; Supplementary Table S1). Nevertheless, we must stress that the primary focus of NASP is the identification of base pairings with a false positive rate that is as low as possible: a focus that should prove particularly useful in studies aiming to evaluate the function of evolutionarily conserved (and therefore probably functional) nucleic acid secondary structures in that it should substantially reduce the time and expense needed to home in on those structures with the greatest biological relevance.

ACKNOWLEDGEMENTS

R. Duffett, G. Botta and A. Meintjes for their technical assistance.

Funding: The Claude Leon Foundation and Wellcome Trust (grant number GR079127MA).

Conflict of Interest: none declared.

REFERENCES

Babak,T. *et al*., (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, **8**, 33.

Bernhart,S.H. *et al*. (2008) RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Gardner,P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140–157.

Hamada,M. *et al*. (2009) Predictions of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25** 465–473.

Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*., **31**, 3423–3428.

Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Method Mol Bio.*, **453**, 3–31.

Pedersen,J.S. *et al*. (2006) Identification and classification of secondary structures in the human genome. *PLoS Comput. Bio*., **2**, e33.

Simmonds,P. *et al*. (2004) Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for evolution and host persistence. *RNA*, **10**, 1337–1351.