RNA 2.5D Structure Prediction ~~using Correlated Substitutions~~

Hamilton, R.S.[1,*], Sadowski, M.I.[2], Davis, I[1], and Taylor, W.R.[2]

[1] *Department of Biochemistry, South Parks Road, University of Oxford, Oxford, OX1 3QU*
[1] *Division of Mathematical Biology, MRC National Institute for Medical Research, London, NW7 1AA*
[*] *Correspondence to: Russell.Hamilton@bioch.ox.ac.uk*

Abstract

RNAs, from the moment they are transcribed, are involved in many diverse processes such as pre-mRNA splicing, polyadenylation, mRNA nuclear export, transport, anchoring, translation and degradation. At the heart of these processes are interactions with the RNAs secondary and tertiary structure with RNA-binding proteins and ncRNAs. The loss or gain of these interactions have been implicated in many human diseases (Cooper *et al.*, 2009). Therefore the accurate prediction of RNA structure is of great importance (Hamilton *et al.*, 2013). Minimum free energy (MFE) methods such as UNAfold (Zuker, 2000) are limited to single sequence secondary structure and canonical base pairings (A-U,G-C,G-U). More recently the use of multiple sequence alignments has improved secondary structure prediction accuracy (Gardner and Giegerich, 2004). Stochastic context-free grammars (SCFG) methods are able to encode all combinations of base pairings and are particularly suited to RNA secondary structure searches (Nawrocki *et al.*, 2009). However each nucleotide can only participate in one base-pairing, ruling out the prediction of pseudoknots, base-pair triplets and more complex motifs such as the kink-turn.

By observing the conservation and any covariation between an RNA and interacting partner (itself, RNA or protein) determines the RNA interaction sites as both the RNA and partner are under the same evolutionary pressure to maintain specific binding for their biological function. We are utilising correlated mutations, initially developed for protein structure contact predicion (Taylor *et al.*, 2012), to predict the 2D and 2.5D contacts with in an RNA. These contacts are not limited to single contacts per base or to canonical base pairings. The predicted contacts will then be used to model the 3D structures of the RNAs giving valuable insight into the RNA surfaces available for interactions. The 2.5D method outperforms the current state-of-the-art SCFG methods and provides contacts for tertiary structure prediction (Taylor *et al.*, 2013). With further development we plan to apply correlated mutation analysis to the prediction of RNA:protein interactions providing crucial insight into disease mechanisms by identifying binding consensus sequences.

Introduction

Focus on shorter RNA, predicting the entire length of an mRNA isn't biologically relevant.
Stabilized upon binding to their protein partners (Martina Doetsch, 2011)
Local environment around a potential binding site determines the accessibility and stability of the motif (Lange *et al.*, 2012)
Structural constraints identified with covariation analysis in ribosomal RNA (Shang *et al.*, 2012)
Survey of all possible base pairings (Abu Almakarem *et al.*, 2012) The coarse-grained modelling will be achieved through the adaption of collision detection algorithms (Katsimitsoulia and Taylor, 2010) and using the NAB package of Amber (Pearlman *et al.*, 1995).

Methods

RNA2.5D Capabilities

RNA2.5D is provided by a Python program, `RNA2.5D.py`

1. ✓MI (Taylor and Sadowski, 2011)

$$MI = S_i + S_j - 2S_{ij} \tag{1}$$

where:
S is the entropy of the position i, j or between i and j

$$S_i = -\sum_{a=1}^{N} \frac{f_a \cdot \log f_a}{\log N} \tag{2}$$

where:
N is the alphabet size of 5 [A,U,C,G,-]
f_a is the frequency of nucleotide a

$$f_a = \frac{(\frac{1}{N} + c_a)}{n+1} \tag{3}$$

where:
c_a is the count of nucleotide of type a over n sequences for each column. This provides a pseudocount to avoid a $log(0)$ if a particular nucleotide is absent.

$$S_{ij} = -\sum_{a=1}^{N}\sum_{b=1}^{N} \frac{f_{ab} \cdot \log f_{ab}}{\log N^2} \tag{4}$$

2. ✓MIp (Normalised MI) (Dunn *et al.*, 2008)

$$MIp = MI - \text{APC} \tag{5}$$

$$APC_{ij} = \frac{MI_{i,\bar{x}} \cdot MI_{j,\bar{x}}}{\bar{MI}} \tag{6}$$

where:
Average product correction
$MI_{i,\bar{x}}$ is the average MI for position i

3. • MIs (MIp with stacking interactions) (Kreth and Fodor, 2014) and (Lindgreen *et al.*, 2006)
$$MIs = MIp + ??? \tag{7}$$

4. • DI (Sadowski *et al.*, 2011; Jones *et al.*, 2012)
$$Dij = \text{inverse of MIs} \tag{8}$$

   Moore-Penrose inversion
   Graphical Lasso

5. ✓DI with MFE enhancement Using RNAfold with constraints to fill in any missing contacts

Datasets and Processing of the data

Datasets RFAM v11 (Griffiths-Jones, 2003)
RFAM families – with known structures (60 families)

Extraction of contacts from PDB Structures

RNAView (Yang *et al.*, 2003) is run for all RFAM with structure families with `AnalyseStructures.pl`

Some PDB files were removed from the analysis as RNAView failed to extract any RNA contact information.
1giy, 1p86, 2rdo, 4adx, 2zkr, 1tf6, 2dlc, 4ari, 486d, 3fic

Estimation of expected number of contacts

Perl script, `CalcAveContactsInRFAM.pl`, calculates the number of contacts in each PDB file linked to RFAM.

Comparison of secondary structure prediction to true 3D

Perl script, `PDB_vs_RFAM.pl`, to convert coordinates between 2D RFAM alignment and 3D PDB structure.

Competing methods to compare to RNA2.5D

Other comparison papers:
(Freyhult *et al.*, 2005)
(Gardner and Giegerich, 2004)
Which datasets did they use for the benchmarking?
ROC plots

Correlated Substitution Based Methods

1. MI and MIp (Dunn *et al.*, 2008) and (Taylor and Sadowski, 2011)
2. di7 (Sadowski *et al.*, 2011)
3. PSICOV Jones *et al.* (2012)
4. DCA Morcos *et al.* (2011)

Minimum Free Energy and Stochastic Context-Free Grammar Based Methods

1. RNAfold (Lorenz *et al.*, 2011)
2. Mfold (Zuker, 2000)
3. Infernal Nawrocki *et al.* (2009)
4. Sfold (Ding and Lawrence, 2003),
5. Pseudoknot capable (hotknots, pknots)
6. 2.5D capable rmdetect
7. Freyhult RNA prediction using MI - email for code

Using predicted contacts to model 3D structure

Proof of concept:
Using NAB with constraints from predictions, followed by a minimisation step (e.g. Amber).

Take a handful of examples (2gis etc) and use predicted contacts to recapitulate 3D.

1. Mfold 2D
2. SCFG 2D
3. RNA2.5D (2.5D contacts at selection of cut offs)

Comparison (RMSD) against known 3D. Aim to determine required number of contacts to predict accurate 3D.

Figure: Panel A. plot number of contacts vs RMSD. Panel B. Overlay of structures?

Results

Worked Prediction Example

In addition to highlighting 2GIS.pdb as a prediction example, RNA2.5D will be used on a real world example (GRIK4). There is an entry in RFAM (RF01383) but no 3D structure. It will follow on from a 2D prediction I performed in 2009 (Pickard *et al.*, 2008). I think there is a possible kink-turn motif in the structure, which is key to it's function. The KT motif is calcium sensitive, in high Ca^{2+} a 60° kink is introduced in to the structure.

Accn	ID; Type	Seqs	PDB	DI7	MIP	DCA	psicov
RF00001	5S_rRNA; Gene; rRNA;	229497	268	✓	•	•	•
RF00002	5_8S_rRNA; Gene; rRNA;	375612	298	✓	•	•	•
RF00003	U1; Gene; snRNA; splicing;	16344	11	✓	✓	✓	•
RF00004	U2; Gene; snRNA; splicing;	11870	1	✓	✓	✓	•
RF00005	tRNA; Gene; tRNA;	298470	440	✓	•	•	✓
RF00010	RNaseP_bact_a; Gene; ribozyme;	6397	6	✓	✓	✓	•
RF00011	RNaseP_bact_b; Gene; ribozyme;	1334	6	✓	✓	✓	✓
RF00015	U4; Gene; snRNA; splicing;	9417	14	✓	✓	✓	•
RF00017	Metazoa_SRP; Gene;	22685	10	✓	✓	✓	•
RF00023	tmRNA; Gene;	5983	26	✓	✓	✓	•
RF00025	Telomerase-cil; Gene;	31	1	✓	✓	✓	•
RF00026	U6; Gene; snRNA; splicing;	72126	1	✓	✓	✓	✓
RF00028	Intron_gpI; Intron;	59999	177	✓	✓	✓	•
RF00029	Intron_gpII; Intron;	51464	10	✓	✓	✓	•
RF00037	IRE_l; Cis-reg;	3104	2	✓	✓	✓	✓
RF00044	Phage_pRNA; Gene;	8	6	✓	✓	✓	•
RF00050	FMN; Cis-reg; riboswitch;	4516	16	✓	✓	✓	✓
RF00059	TPP; Cis-reg; riboswitch;	11197	18	✓	✓	✓	•
RF00061	IRES_HCV; Cis-reg; IRES;	7721	4	✓	✓	✓	•
RF00094	HDV_ribozyme; Gene; ribozyme;	598	14	✓	✓	✓	✓
RF00100	7SK; Gene;	21885	1	•	✓	✓	•
RF00114	S15; Cis-reg; leader;	1082	2	✓	✓	✓	✓
RF00162	SAM; Cis-reg; riboswitch;	4757	20	✓	✓	✓	✓
RF00163	Hammerhead_1; Gene; ribozyme;	49036	2	✓	✓	✓	•
RF00164	s2m; Cis-reg;	641	1	✓	✓	✓	✓
RF00167	Purine; Cis-reg; riboswitch;	2427	26	✓	✓	✓	✓
RF00168	Lysine; Cis-reg; riboswitch;	2422	14	✓	✓	✓	✓
RF00169	Bacteria_small_SRP; Gene;	5622	4	✓	✓	✓	✓
RF00175	HIV-1_DIS; Cis-reg;	3631	44	✓	✓	✓	✓
RF00177	SSU_rRNA_bacteria; Gene; rRNA;	7429	436	✓	✓	✓	•
RF00207	K10_TLS; Cis-reg;	15	3	✓	✓	✓	•
RF00220	Rhino_CRE; Cis-reg;	122	1	✓	✓	✓	✓
RF00234	glmS; Cis-reg; riboswitch;	842	37	✓	✓	✓	✓
RF00252	Alfamo_CPB; Cis-reg;	45	1	✓	✓	✓	•
RF00261	IRES_L-myc; Cis-reg; IRES;	141	2	✓	✓	✓	✓
RF00374	Gammaretro_CES; Cis-reg;	1356	4	✓	✓	✓	✓
RF00380	ykoK; Cis-reg;	1493	3	✓	✓	✓	✓
RF00436	UnaL2; Cis-reg;	100501	1	✓	•	✓	•
RF00458	IRES_Cripavirus; Cis-reg; IRES;	24	7	✓	✓	✓	•
RF00480	HIV_FE; Cis-reg; frameshift_element;	25285	3	✓	✓	✓	✓
RF00500	TCV_H5; Cis-reg;	5	1	✓	✓	✓	•
RF00504	Glycine; Cis-reg; riboswitch;	6875	21	✓	✓	✓	✓
RF00522	PreQ1; Cis-reg; riboswitch;	894	5	✓	✓	✓	✓
RF00524	R2_retro_el; Cis-reg;	802	28	✓	✓	✓	✓
RF00618	U4atac; Gene; snRNA; splicing;	715	4	✓	✓	✓	✓
RF01051	c-di-GMP-I; Cis-reg;	1990	19	✓	✓	✓	✓
RF01073	GP_knot1; Cis-reg;	7835	1	✓	✓	✓	✓
RF01118	PK-G12rRNA; Gene; rRNA;	23118	100	✓	✓	✓	✓
RF01510	MFR; Cis-reg; riboswitch;	4	16	✓	✓	✓	•
RF01734	crcB; Cis-reg;	1267	5	✓	✓	✓	✓
RF01786	c-di-GMP-II; Cis-reg; riboswitch;	237	2	✓	✓	✓	✓
RF01831	THF; Cis-reg; riboswitch;	598	7	✓	✓	✓	✓
RF01852	tRNA-Sec; Gene; tRNA;	1959	6	✓	✓	✓	✓
RF01854	Bacteria_large_SRP; Gene;	1662	2	✓	✓	✓	✓
RF01857	Archaea_SRP; Gene;	275	11	✓	✓	✓	✓
RF01959	SSU_rRNA_archaea; Gene; rRNA;	7394	85	✓	✓	✓	•
RF01960	SSU_rRNA_eukarya; Gene; rRNA;	425	69	✓	✓	✓	•
RF01998	group-II-D1D4-1; Intron;	1726	1	✓	✓	✓	✓
RF02001	group-II-D1D4-3; Intron;	2450	6	✓	✓	✓	✓
RF02095	mir-2985-2; Gene; miRNA;	302	1	✓	✓	✓	✓

Table 1: RFAM Results for all the methods

Discussion

Where correlated mutations do well
Where correlated mutations don't do well
Hybrid method correlated mutations + MFE
Ultimate goal is RNA:protein
How to pair up RNA seqs with protein seqs

References

Abu Almakarem,A.S., Petrov,A.I., Stombaugh,J., Zirbel,C.L. and Leontis,N.B. (2012) Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic acids research,* **40** (4), 1407–1423.

Cooper,T.A., Wan,L. and Dreyfuss,G. (2009) RNA and Disease. *Cell,* **136** (4), 777–793.

Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic acids research,* **31** (24), 7280–7301.

Dunn,S.D., Wahl,L.M. and Gloor,G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics,* **24** (3), 333–340.

Freyhult,E., Gardner,P.P. and Moulton,V. (2005) A comparison of RNA folding measures. *BMC Bioinformatics,* **6**, 241.

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

Griffiths-Jones,S. (2003) Rfam: an RNA family database. *Nucleic acids research,* **31** (1), 439–441.

Hamilton,R.S., Ball,G. and Davis,I. (2013) A Multidisciplinary Approach to RNA Localisation. In *Biophysical approaches to translational control of gene expression*. Springer pp. 213–233.

Jones,D.T., Buchan,D.W.A., Cozzetto,D. and Pontil,M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics,* **28** (2), 184–190.

Katsimitsoulia,Z. and Taylor,W.R. (2010) A hierarchic collision detection algorithm for simple Brownian dynamics. *Computational Biology and Chemistry,* **34** (2), 71–79.

Kreth,K.E. and Fodor,A.A. (2014) Covariance in protein multiple sequence alignments using groups of columns. *arXiv preprint arXiv:1401.1141,* **arXiv**.

Lange,S.J., Maticzka,D., Möhl,M., Gagnon,J.N., Brown,C.M. and Backofen,R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic acids research,* **40** (12), 5215–5226.

Lindgreen,S., Gardner,P.P. and Krogh,A. (2006) Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics,* **22** (24), 2988–2995.

Lorenz,R., Bernhart,S.H., Hoener zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms for Molecular Biology,* **6** (1), 26.

Martina Doetsch,R.S.B.F. (2011) Transient RNA–protein interactions in RNA folding. *The FEBS journal,* **278** (10), 1634.

Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences,* **108** (49), E1293–301.

Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics,* **25** (10), 1335–1337.

Pearlman,D., Case,D.A., Caldwell,J., Ross,W., Cheatham,T., DeBolt,S., Ferguson,D., Seibel,G. and Kollman,P. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications,* **91** (1), 1–41.

Pickard,B.S., Knight,H.M., Hamilton,R.S., Soares,D.C., Walker,R., Boyd,J.K.F., Machell,J., Maclean,A., McGhee,K.A., Condie,A., Porteous,D.J., St Clair,D., Davis,I., Blackwood,D.H.R. and Muir,W.J. (2008) A common variant in the 3'UTR of the GRIK4 glutamate receptor gene affects transcript abundance and protects against bipolar disorder. *Proceedings of the National Academy of Sciences,* **105** (39), 14940–14945.

Sadowski,M.I., Maksimiak,K. and Taylor,W.R. (2011) Direct correlation analysis improves fold recognition. *Computational Biology and Chemistry,* **35** (5), 323–332.

Shang,L., Xu,W., Ozer,S. and Gutell,R.R. (2012) Structural Constraints Identified with Covariation Analysis in Ribosomal RNA. *PLoS ONE,* **7** (6), e39383.

Taylor,W.R., Hamilton,R.S. and Sadowski,M.I. (2013) Prediction of contacts from correlated sequence substitutions. *Current Opinion in Structural Biology,* **23** (3), 473–479.

Taylor,W.R., Jones,D.T. and Sadowski,M.I. (2012) Protein topology from predicted residue contacts. *Protein science : a publication of the Protein Society,* **21** (2), 299–305.

Taylor,W.R. and Sadowski,M.I. (2011) Structural Constraints on the Covariance Matrix Derived from Multiple Aligned Protein Sequences. *PLoS ONE,* **6** (12), e28265.

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 research,* **31** (13), 3450–3460.

Zuker,M. (2000) Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology,* **10** (3), 303–310.