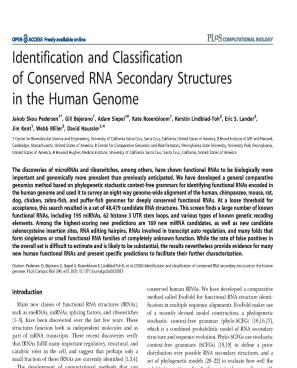
# Automating the Classification of

Authorship & Acknowledgement

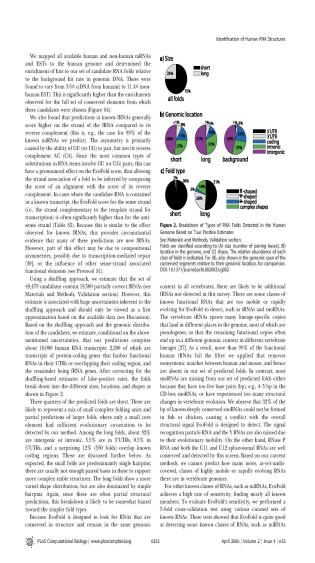
NIC Weber nmweber@illinois.edu Andrea Thomer akthom2@illinois.edu @\_an\_dre\_a

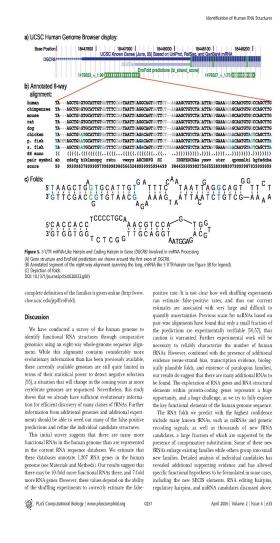
## Motivation

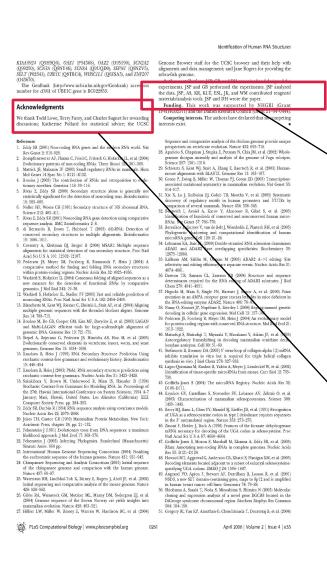
- > Explore the automation of classifying acknowledgment & authorship
- > Bootstrap the use of existing ontologies to increase the reliability of our own classifications



ntroduction	conserved human fRNAs. We have developed a comparative method called EvoFold for functional RNA-structure identi-
Many new classes of functional RNA structures (RNAs), ach as snoRNAs, miRNAs splicing factors, and ribosotiches 15-3, have been discovered over the last few years. These tructures function both as independent molecules and as sart of mRNA transcripts. These recent discoveries verify hat (RNAs falfill many important regulators, structural, and adaptive roles in the cell and suggest that perhaps only a malf fraction of these (RNAs are currently identified [1,54]. The development of computational methods that can efficiently identify fixNAs by comparative genomics has been nameered by the fact that RNAs often exhibit only weakly conserved primary-sequences signals [5]. Fortunately, the	fication in multiple sequence alignments. EvoFold makes use of a recently devised model construction, a phylogenetic stochastic context-free grammar [phylos-CFG [0],61,67], which is a combined probabilistic model of RNA secondary structure and sequence evolution. Phylos-CFG us us tochastic context-free grammars (SCFG) [18,19] to define a prior distribution over possible RNA secondary structures, and a set of phylogenetic models [20–22] to evaluate how well the substitution pattern of each alignment column conforms with its secondary-structure amountain. EvoFold uses a very general model of RNA secondary structures that allows it to
tem-pairing regions of fRNA structures evolve mostly with a characteristic substitution pattern such that only substitu-	Editor: Richard Durbin, Sanger Institute, United Kingdom
ions that maintain the pairing capability between paired	Received September 8, 2005; Accepted March 6, 2006; Published April 21, 2006
bases will be allowed. This leads to compensatory double substitutions (e.g., GC $\leftrightarrow$ AU) and to a few types of	A previous version of this article appeared as an Early Online Release on March 6, 2006 (DOL 10.1371/journal.pcbi.0020033.eor).
compatible single substitutions (e.g., GC $\leftrightarrow$ GU); the latter	DOI: 10.1371/journal.pcbi.0020033
nade possible by RNA's ability to form a non-Watson-Crick oair between G and U. This evolutionary signal can be exploited for comparative identification of fRNAs [6-12]. The many non-human vertebrate genomes now sequenced	Copyright: © 2006 Pedersen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
can be aligned against the human genome, leading to a multiple alignment with considerable information about the evolutionary process at every position [13–15]. Given a	Abbreviations: ADAR (adenosine deaminase acting on RNA; bp, base pair; DGCR8, DiGeorge syndrome critical region; fps, folding potential score; RNA, functional RNA structures, Phylo-SFCE, phylogenetic stochastic context free grammar; SEOS, selenocysteine insertion sequence; UCSC, University of California Santa Cruz
liverse enough set of genomes, comparative methods that	* To whom correspondence should be addressed. E-mail: jsp@soe.ucsc.edu
can make effective use of this evolutionary information should in principle be able to efficiently identify the	<sup>II</sup> Current address: Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America
PLoS Computational Biology   www.ploscompbiol.org 025	April 2006   Volume 2   Issue 4   e33







### Data

- > Corpus of articles from the field of Bioinformatics (n= 9741)
- > Extracted authorship statements and acknowledgments (see below) for each article
- Manually classified a subset (n = 300) of each paratext using the Scholarly Contributions and Roles Ontology (Shotton and Peroni, 2013)

### Authorship

Author contributions. JSP, GB, and DH conceived and designed the experiments. JSP and GB performed the experiments. JSP analyzed the data. JSP, AS, KR, KLT, ESL, JK, and WM contributed reagents/materials/analysis tools. JSP and DH wrote the paper.

#### Acknowledgement

We thank Todd Lowe, Terry Furey, and Charles Sugnet for rewarding discussions; Katherine Pollard for statistical advice; the UCSC Genome Browser staff for the UCSC browser and their help with alignments and data management; and Jane Rogers for providing the zebra-fish genome.

## Automation

Using our manual classifications as training data, we attempted to use Stanford's etcML to automate the classifications of each

Full results are available at

http://dx.doi.org/10.6084/m9.figshare. 928642









**Works Cited**