

# RNA Informatics

(Europe's biggest cycle-user?):  
what we're doing about it/  
what we're learning

W.L. Ruzzo

UW CSE

11/04/2004



## Rfam: an RNA database

- 1/2003: rel. 1.0 - 36 entries
- 6/2004: rel. 6.1 - 379 entries

Biggest scientific computing  
user in Europe ---  
1000 cpu's for a month per rel

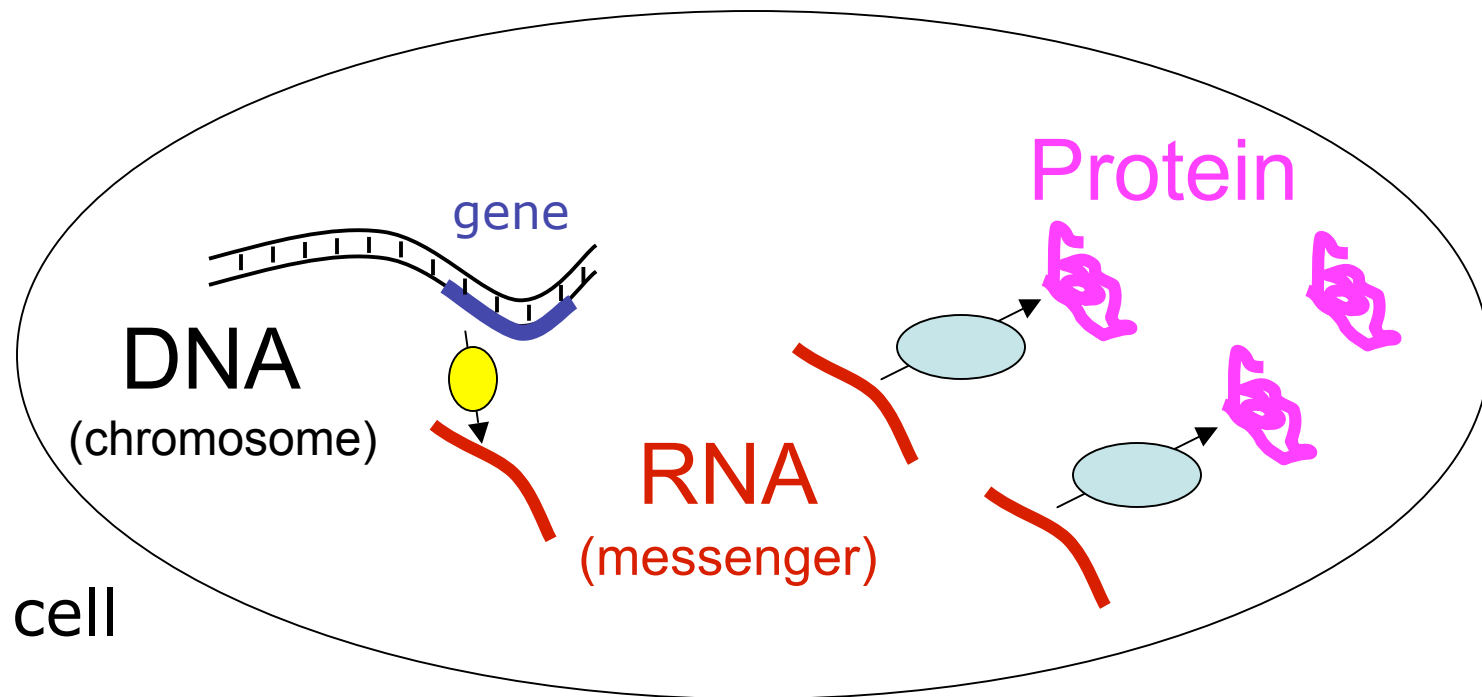
What We're  
Doing:

Making it Slower

**And Proud of  
it!**

# The “Central Dogma”

DNA → RNA → Protein

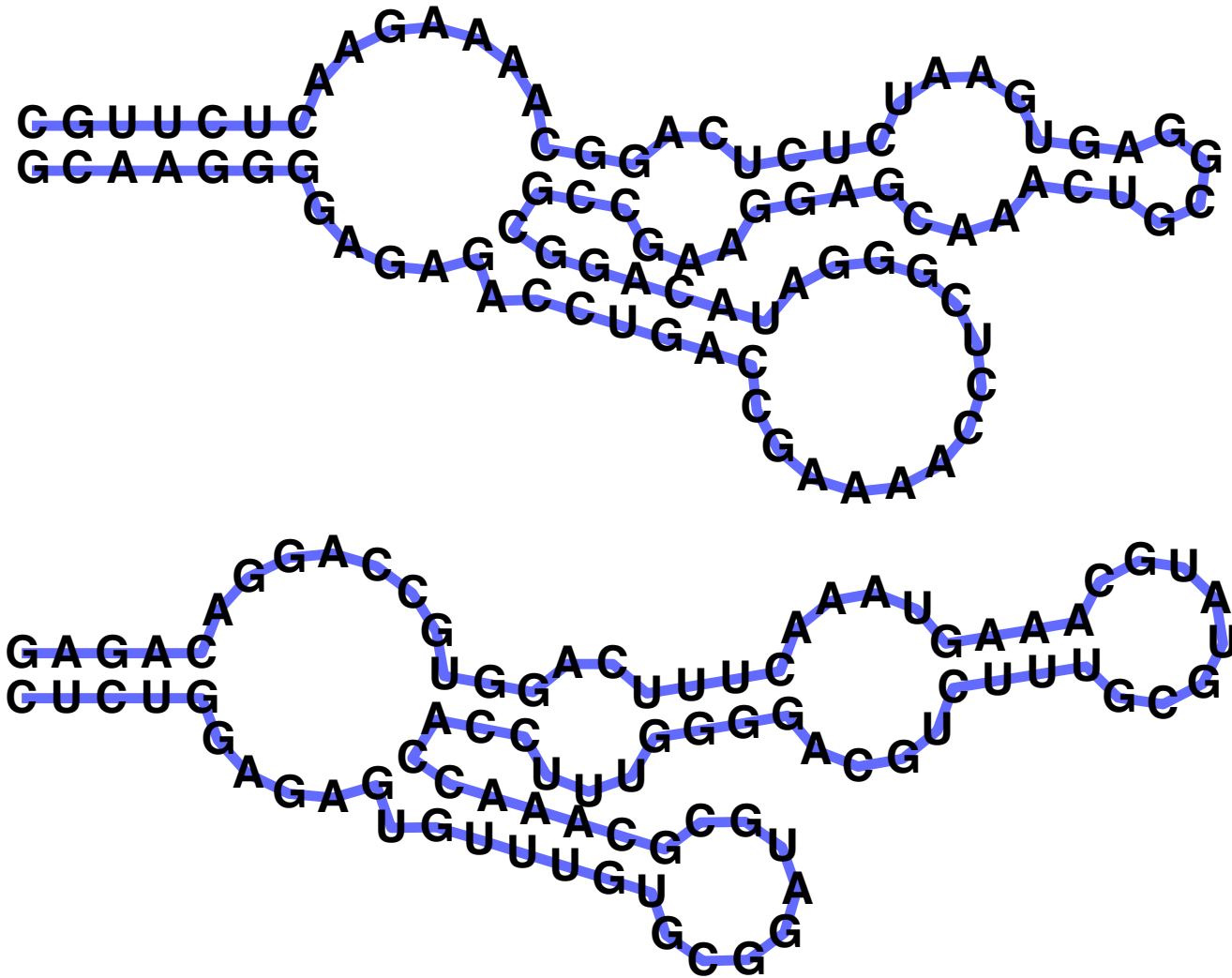


# A Genome Project Surprise: Non-coding RNA

- Messenger RNA - codes for proteins
- Non-coding RNA - all the rest
  - Before, say, mid 1990's, 1-2 dozen known (critically important, but narrow roles: e.g. ribosomal and transfer RNA, splicing, SRP)
- Since mid 90's dramatic discoveries
  - Regulation, transport, stability/degradation
  - E.g. "microRNA":  $\approx$  250 in humans



# Q: What's so hard?

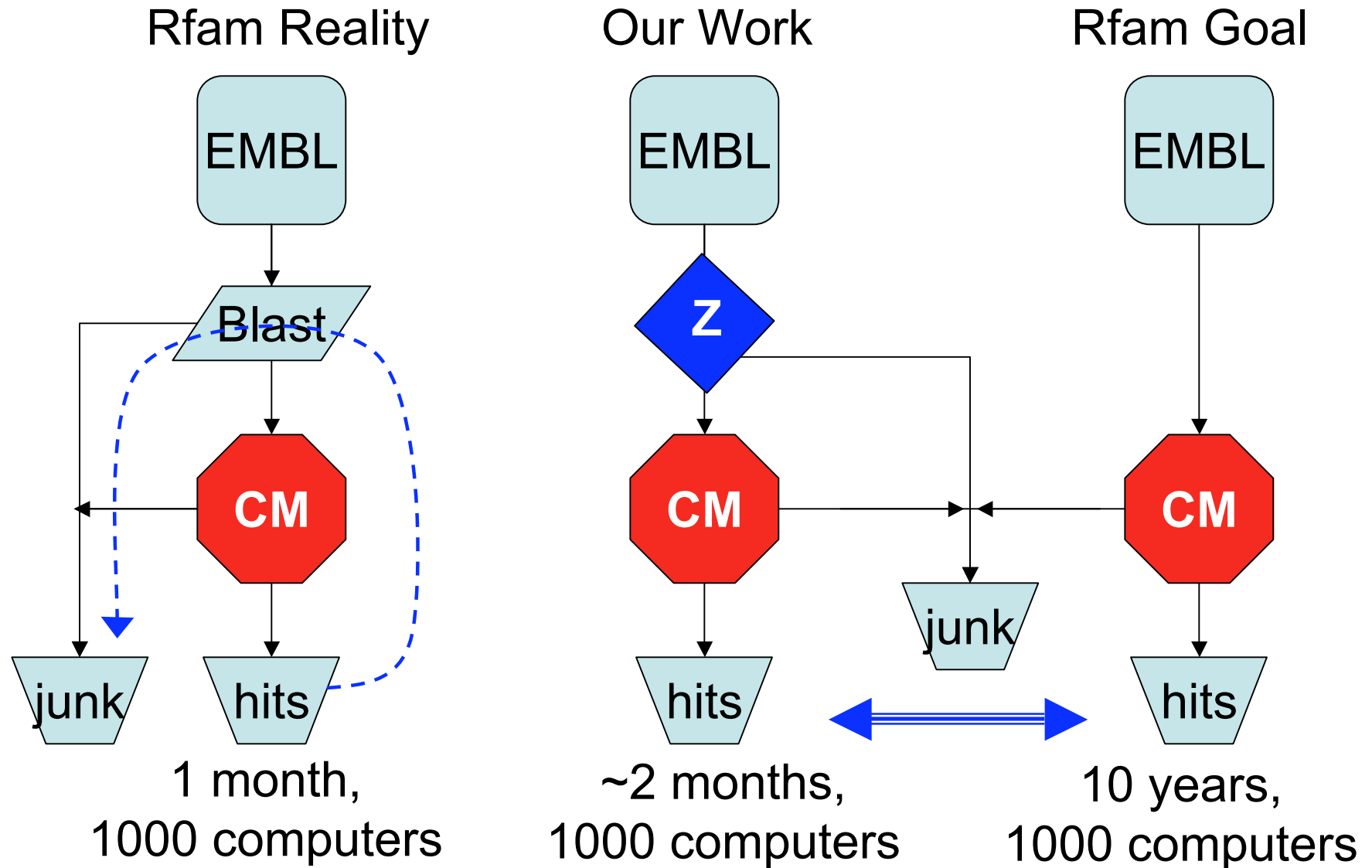


A: Structure often more important than sequence

# Computational Challenges

- Challenge 1:  
Better search

# What we're doing about it




## Results: Many new Finds

	# with BLAST + CM	# with rigorous filter series + CM	# new
Rfam tRNA	58609	63767	5158
Group II intron	5708	6039	331
tRNAscan-SE (human)	608	729	121
<i>Pyrococcus</i> snoRNA	57	180	123
Iron response element	201	322	121
Histone 3' element	1004	1106	102
Purine riboswitch	69	123	54
Lysine riboswitch	60	71	11
Hammerhead I	167	193	26
And more...			

# Covariance Models

(specialized stochastic CFGs)

Sequences

CAG or AAU  




CM

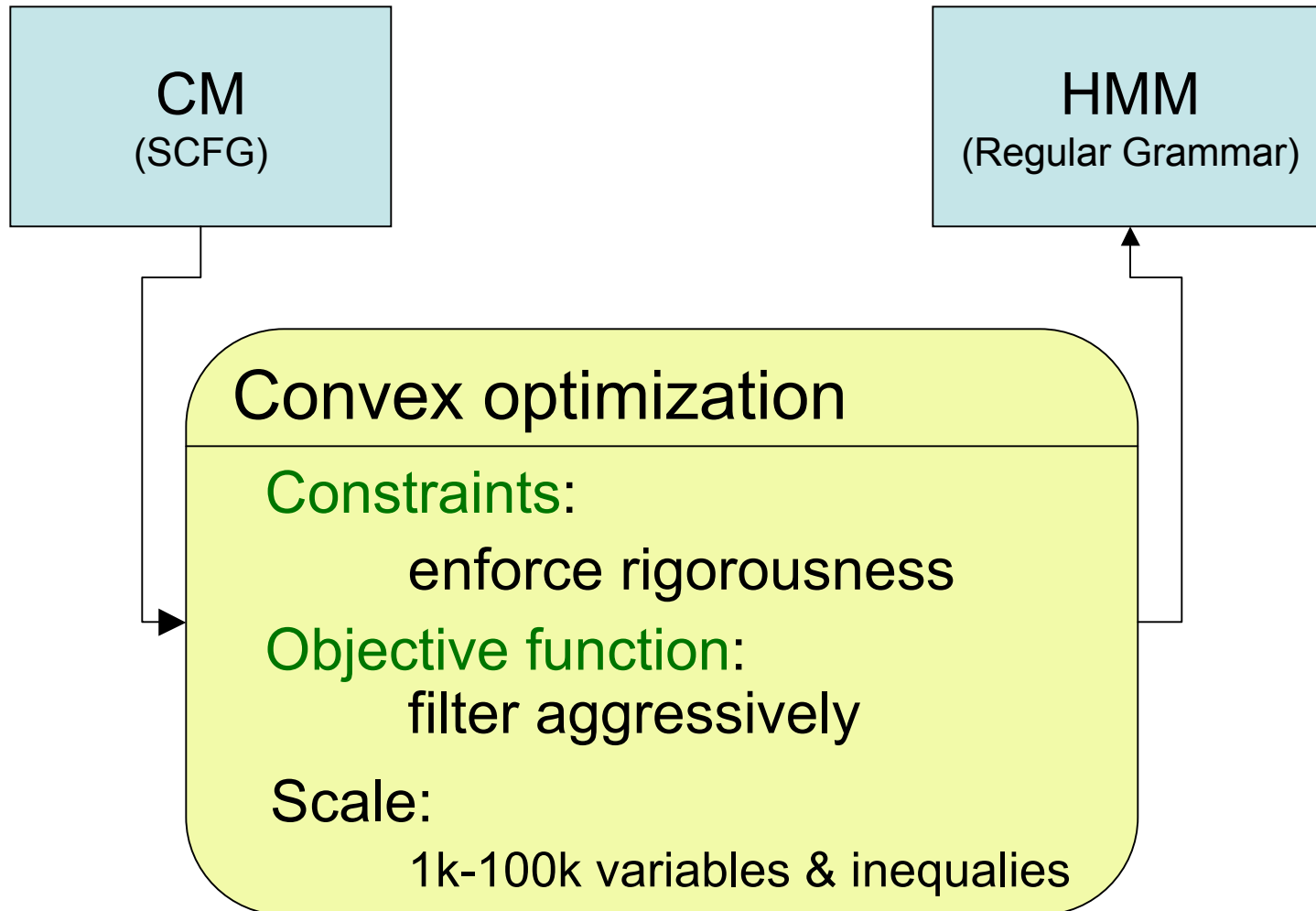
$S_1 \rightarrow cS_2g \mid aS_2u$

$S_2 \rightarrow a$

Example  
parse of CAG

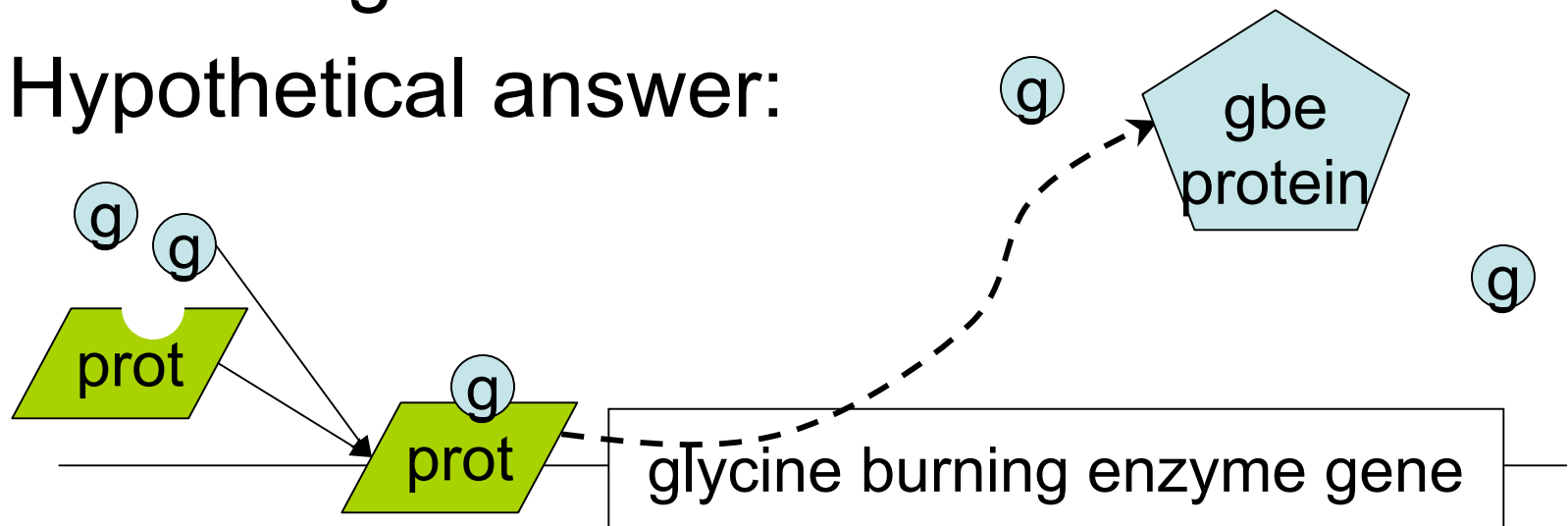
$S_1 \rightarrow cS_2g \rightarrow cag$

# How to go faster (carefully)



# Another Example: the Glycine Riboswitch

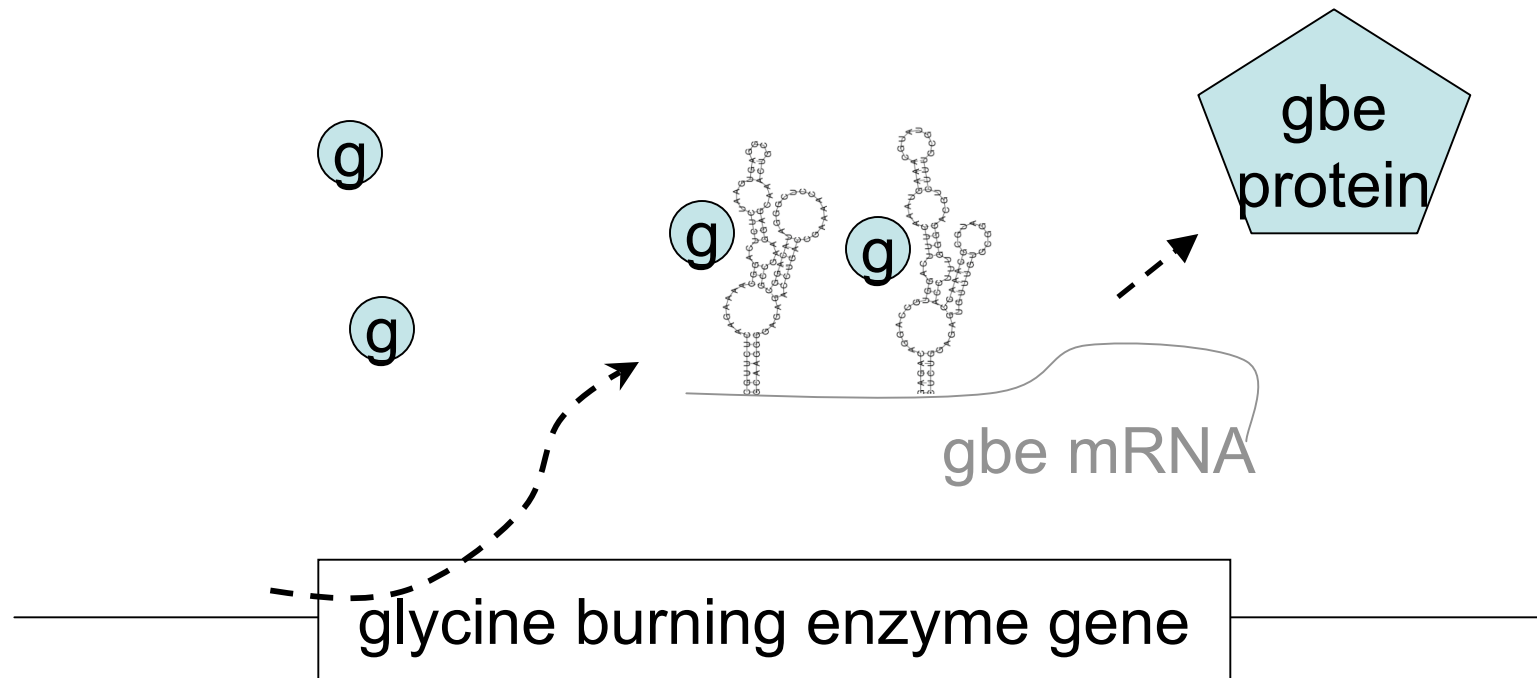
- Glycine - simplest amino acid
- Uses - make proteins, make energy
- Not enough OR too much - wasteful
- Hypothetical answer:



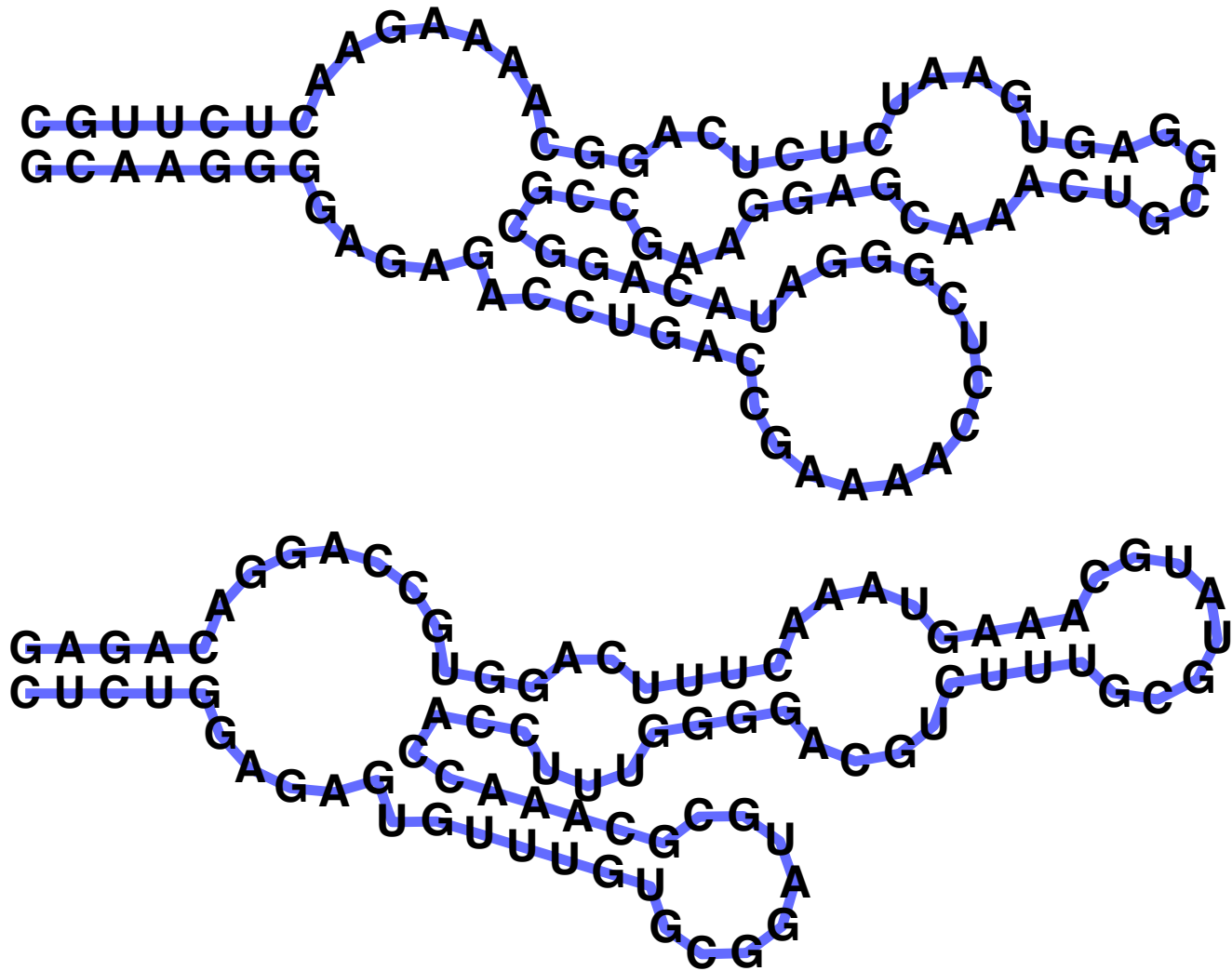
# The Glycine Riboswitch

Mandal, Lee, Barrick,  
Weinberg, Emilsson,  
Ruzzo, and Breaker,  
*Science*, 10/2004

- Actual answer (in many bacteria):  
Look Ma, no protein



# The Glycine Riboswitch

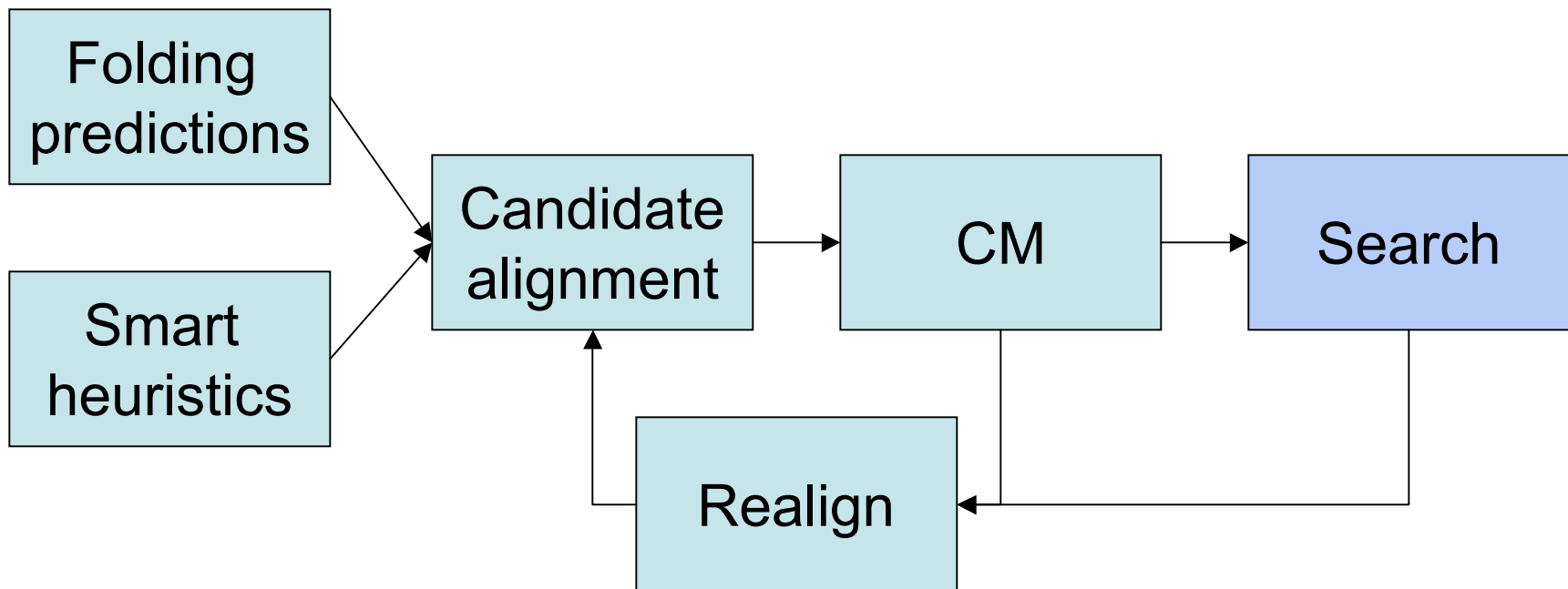


# Computational Challenges

- Challenge 1:  
Better search
- Challenge 2:  
Better model construction/discovery

# CMFinder

- Finding CM models *without* alignment

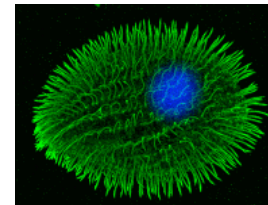
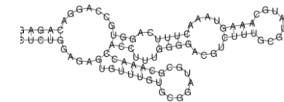


# CMFinder Example

- Start with 12 related bacterial genes from Footprinter (Tompa et al.)
- Found 9 with good pattern
- Searched all bacterial genomes
- Found 234 hits
- Realigned them
- Result is about 90% similar to Rfam T-box
  - (Based on hand-curated alignment of 67 knowns)

# Summary

- ncRNA bioinformatics is important, hard
- “Rigorous Filtering” :  
100-fold speedup, *no* loss of accuracy
- Automated model discovery attainable
- Application to riboswitch  
discovery, genome annotation  
in progress



# Acknowledgements

Zasha Weinberg

Zizhen Yao

Katarzyna Wilamowska

Martin Tompa & students

Ron Breaker, Jeff Barrick (Yale)