Cocke-Kasami-Younger Parser

Suppose all rules of form $A \rightarrow BC$ or $A \rightarrow a$ (by mechanically transforming grammar)

Given
$$x = x_1 \dots x_n$$
, want $M_{i,j} = \{ A \mid A \Rightarrow^* x_{i+1} \dots x_j \}$

For j=2 to n $M[j-1,j] = \{A \mid A \rightarrow x_{j} \text{ is a rule} \}$ for i = j-1 down to 1 $M[i,j] = \bigcup_{i < k < j} M[i,k] \otimes M[k,j]$ Where X \otimes Y = {A | A \rightarrow BC , B \in X, and C \in Y }



Time: O(n³)

And now for something completely different

CFGs beyond compilers

An RNA Structure



An RNA Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off

A mRNA leader

		TCC			P1				
	-35 -10	135			P2	_		885	Start
-		40.0000							
290	10000000 I7 (0000000)	TOTHER CORD	souceseucu	Gees G	000006	OSGC. CAL	GREEKOLD	510510sG66	108.405
3ñe	103016.17.200101.	. 17 . AUUACCAD	suncese <mark>ne</mark> .		GGURGANG	. COGUCAU	GAGGAUGD	S. CO. REGIS	.11.405
oik	TEGRAC, 17 . SAMASI	31.UARAC <mark>GRO</mark>	soucese <mark>us</mark> .	020	CAUACUU	GUUCAU	GAGCAUU	G.06.RECAC	.07. <mark>RUG</mark>
Bee	TEGOTA, 18, PARGOT.	36.UURAC <mark>GRU</mark>	suucese <mark>us</mark> .	URA . OUCU	AUURAGACU.	ARU AUUR.	GAGCAUC	G.05. <mark>RGCLC</mark>	.09. <mark>806</mark>
Gita	rescer.17. Harcar.	38. AARAC <mark>GRU</mark>	suuc <mark>ese<mark>us</mark>.</mark>	CRAUGA.	AGAGAUC	auus seau	GARCAUC	G.01. AGGAG	.08. <mark>AUG</mark>
Bc2	TECEC. 17. PARCAT.	45.AUUAC <mark>GAU</mark>	AUUCCCCC <mark>UC</mark> .	CUG	CAGUGU	.UGG.CAU	GAAUGUCU	0.06. <mark>40040</mark>	.10.AUG
Sec.	APGACA.17. GARAGE.	35. AURAC	soucese <mark>us</mark> .	CR. RUAR	AGARAGUEUG	UG. CAA	GAGCAUC	G. 05. RECAC	.08.AUG
$L\pi\phi$	TPIACA, 17. PAACOL	28. AUAAC <mark>CAU</mark>	AUUCCECC <mark>UU</mark> .	CAU	UAUUAAU	. AUG. AAU	CAAUCUU	0.05. <mark>ROCAC</mark>	.07. <u>AUC</u>
Sau	терала, 17. сласат.	23.AUCAC <mark>OAD</mark>	saucese <mark>us</mark> .	cuauai	URUUUGUCG.	AGGCAA	GARCAUR	5.04. <mark>RGRGS</mark>	.09. <mark>805</mark>
Cpe	TTAAA0, 18. TAAACI.	08.0UACC	aeuceue <mark>ue</mark> u	CACA	.080	USUSUUA.A	onacouca.	A.17. <mark>200200</mark>	.08.203
Chy	TESCAR, 17. PARAAT.	09.UACCRARCO	soucese <mark>us</mark> .	GR CI	RGGGGGC	UC.CAU	GARCOUGO	c.03.RGGAG	.09. <mark>806</mark>
880	TTGAGA, 17 . TRABAT,	16.RARARCON	egucese <mark>us</mark> .	CRUU AL	AACUAR	ARUG. URU	GRACACCU	0.05. <mark>RCCAC</mark>	.07. <mark>AUG</mark>
Ante	TTOOOG.17. TATAXT.	10.UUACGGOC	souceue <mark>ura</mark> .	URC	. AGGA	GUA. UAA	GAACOUC <mark>U</mark>	N.07. <mark>RGGRO</mark> G	.07.803
DTO	TPOCCO, 17. TARANI.	16.UUACC <mark>CAC</mark>	ceuccee <mark>uc</mark> .	CCU	CUCCCAA	ACC. URA	CARCOUCU.	a.01. accan	.12. <mark>CUC</mark>
Spa	TTTRET. 17. PARKET.	. 28 . AUACA <mark>CIUU</mark> 1	u <mark>auc</mark> ese <mark>us</mark> .	AGGA	.agau	DCCU. CAA	GAUGGACA	A.04. RGGROJ	.05.803
Sau	TTTACA.17. TACANT.	. 26 . ЛАЛСС <mark>ССШ</mark>	AAUCCECC <mark>UC</mark> .	80 M	CAGAGCA	<mark>CU.UA</mark> U	CAUUACUA	A.01. ACCAO	.07.200
Lp1	TTOCOL, 18. TATTCL.	21.UUAACCAU	auuceae <mark>ue</mark> .	80 · · · · · · ·	CAGOUU	OU. CAC	OVYANOOCO	3.01. <mark>ROCAN</mark>	.09.208
Bfa	TTIACA, 17. TAAACI.	28. AUUACAAU	AUUCCECC <mark>UG</mark> .	096.05.	. 0.5.58 <mark>00</mark>	ACCA. UAA	ολλυλυψο	9.06. <mark>20020/</mark>	.08.403
L_{fo}	TTTACA. 17. TAXACT.	25.UUAUGGOU	AUTUCCOC <mark>ID</mark> O.	GCAC	.AAG	CTUCTUDA AU	oAADOCCO	U.03. <mark>80080</mark> /	.07.803
Sth	TAGACA <mark>. 17.</mark> TAAGAT.	29.UAACGGCU	AAUCCOC <mark>UG</mark> .	AGA . CAC	AGAGGU DO	cucu.uaa	GAUUAGU <mark>A</mark> .	N.03. <mark>RAGAD</mark>	.08.803
Lao	TTARAA. 17. PIACET.	. 39 . UUAUGGOU	AUTUCCOC <mark>U</mark> G.	ACG	CUGGUA	. coupaau	GARDGCC	A.03. <mark>RGGRO</mark>	.10.503
Spy	TTTRCA. 17. PROAMT.	. 29 . JUACG <mark>CCU</mark>	AAUCCGCUA.	XG	CAAGUA	CU.UAA	GAUUAGUA	A.03. <mark>RGGRG</mark>	.06.803
Les	TETERA, 17. TRABAT	26. ACARC	ADUCCESC <mark>U</mark> G.	aca(CARGE	- COUVARU	GAADADOD	9.06. <mark>RGGLG</mark>	.07.803
Le2	HERRON, 17. SACKET.	24.AUAACCEU	ADUCCESC <mark>UC</mark> .	c	AACUG	Gacau	GARDGOCO	9.04. <mark>80634</mark> 0	.07.803
Fac	199300.17. JANANT.	12.AMUUCOAU	AUTUCEGE <mark>UTU</mark> .	CIAR		UUA.AQU	GAADADOU	U.04. RGG3.A	.02.802





An RNA Grammar

- $S \rightarrow LS | L$ $L \rightarrow S | "dFd"$ $F \rightarrow LS | "dFd"$
- "s" means unpaired; "dFd" means paired (Watson–Crick: *aFu* | *uFa* | *gFc* | *cFg* paren-like nesting)

- $S \hspace{.1in} \rightarrow \hspace{.1in} LS \rightarrow LLLLLLLS \rightarrow LLLLLLLL$
 - $\rightarrow ssLsssss \rightarrow ssdFdsssss$
 - \rightarrow ssdddFdddsssss
 - \rightarrow ssdddLSdddsssss
 - \rightarrow ssdddLLLLdddsssss
 - $\rightarrow ssdddssssdddsssss$

$$s^{ss}s$$

 $d-d$
 $d-d$
 $ssd-d_{sssss}$

 $\begin{array}{rcl} F & \rightarrow & dFd \rightarrow ddFdd \rightarrow ddLSdd \\ & \rightarrow & ddLLdd \rightarrow ddLsdd \rightarrow dddFdsdd \end{array}$

Actually, a Stochastic CFG

Associate probabilities with rules, e.g.:

$$S \rightarrow LS$$
 (p = 0.87)
 $S \rightarrow L$ (p = 0.13)

Now we can ask, not only "Does S generate w?" But also "How likely is it?"

Cocke-Kasami-Younger Parser

Suppose all rules of form $A \rightarrow BC$ or $A \rightarrow a$ (by mechanically transforming grammar)

Given
$$x = x_1...x_n$$
, want $M_{i,j} = \{ A \mid A \Rightarrow^* x_{i+1}...x_j \}$

For j=2 to n $M[j-1,j] = \{A \mid A \rightarrow x_{j} \text{ is a rule}\}$ for i = j-1 down to 1 $M[i,j] = \bigcup_{i < k < j} M[i,k] \otimes M[k,j]$ Where X \otimes Y = {A | A \rightarrow BC , B \in X, and C \in Y }



Time: $O(n^3)$

"Inside" Algorithm for SCFG

Suppose all rules of form $A \rightarrow BC$ or $A \rightarrow a$ (by mechanically transforming grammar)

Given
$$x = x_1...x_n$$
, want $M_{i,j}^A = p(A \Rightarrow^* x_{i+1}...x_j)$

For j=2 to n $M^{A}[j-1,j] = p(\text{ rule } A \rightarrow x_{j})$ for i = j-1 down to 1 $M^{A}[i,j] = \sum_{A \rightarrow BC, i < k < j} M^{B}[i,k] \times M^{C}[k,j]$

I.e., *probability* of A in M[i,j], instead of its *possibility*



Time: $O(n^3)$

ncRNA Discovery in Bacteria

Cmfinder--A Covariance Model Based RNA Motif Finding Algorithm, Yao, Weinberg, Ruzzo, *Bioinformatics*, 2006, 22(4): 445-452,

A Computational Pipeline for High Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes. Yao, Barrick,

Weinberg, Neph, Breaker, Tompa and Ruzzo. *PLoS Comput Biol.* 3(7): e126, July 6, 2007.

Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker. *Nucl. Acids Res.,* July 2007 35: 4809-4819.



Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker. *Nucl. Acids Res.,* July 2007

Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa,

ncRNA Discovery in Humans

Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin

Genome Research, Jan '08



cdt. 11

β-actin RT+

control β-actin RT-

control



Bottom Line

CFG technology is a key tool for RNA description, discovery and search

A very active research area. (Some call RNA the

"dark matter" of the genome.)

Huge compute hog: results above represent hundreds of CPU-years, and smart algorithms can have a big impact

More?

Check out CSE 427