The Big ORF Theory:

The Intensive Enumeration, The Triplet Approximation, & The Monte Carlo Simulation

Steven M. Carr
Departments of Biology & Computer Science
Memorial University of Newfoundland

31 March 2026

1

Bioinformatics & "*The Central Dogma*" of DNA

In a **protein-coding gene region**,
Five of six three-letter **reading frames** "***closed***" by "***stop***" triplets
Sixth an "***Open Reading Frame***" (***ORF***) without "*stops*" that specifies **polypeptide**

Bioinformatic "***data mining***" includes mega-scale "***ORF finding***"

"*5&1*" condition can be built into short **dsDNA** exemplars (**L ≤ 25 bp**)
for "***ORF finding***" in genetics & bioinformatic **education**:
"***There's an app for that !***"

Carr's Conjecture: *No* **dsDNA** *of* **L ≤ 10** *exists that satisfies "* **5&1 condition***"*

Simple questions lead to **unexpected answers**
"***5&1 condition***" functions: **enumerations**, **algorithms**, **approximations**, **simulations**
Implications for **ORF** features; Alternative Genetic Codes, etc.

2

Carr's Conjecture

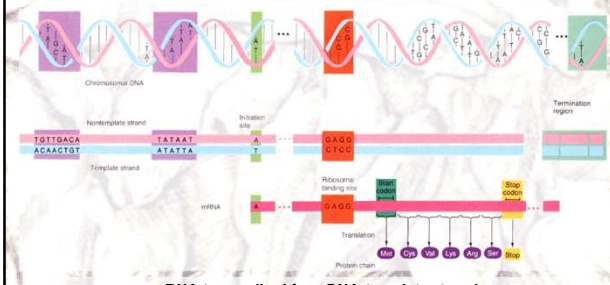
Any protein-coding **DNA** has ***six*** **5' →3' reading frames.**
Five "***closed***" by "***stop***" triplets
Sixth "***Open Reading Frame***" (***ORF***) without "*stops*": specifies **polypeptide**

Carr's Conjecture: *No* **dsDNA** *of* **L ≤ 10** *exists that satisfies the "* **5&1 condition***"*

The 4241 Challenge: Find random **ssDNA** string **L ≥ 11** that satisfies "**5&1 condition"**

3

"Let's Not Make it Harder Than it Has to Be"
DNA chase RNA chase Protein



mRNA transcribed from **DNA *template* strand**
Paired **DNA *sense* strand** has ***same sequence*** as **mRNA** (except **T➔U**)

4

The Genetic Code (1965)

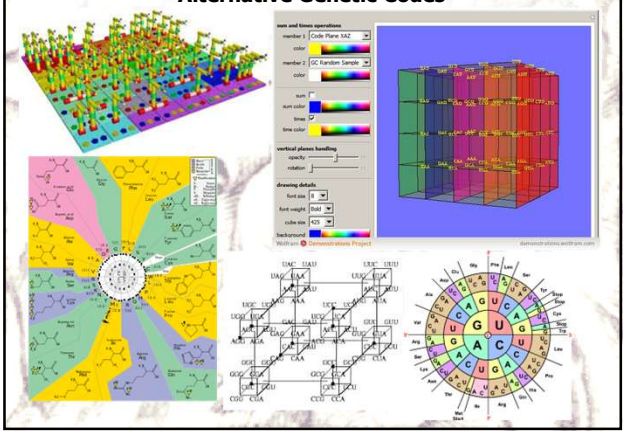
TABLE 3. NUCLEOTIDE SEQUENCES OF RNA CODONS

1st Base	2nd Base U	2nd Base C	2nd Base A	2nd Base G	3rd Base
U	PHE*	SER*	TYR*	CYS*	U
	PHE*	SER*	TYR*	CYS	C
	leu*?	SER	TERM?	cys?	A
	leu*, f-met	SER*	TERM?	TRP*	G
C	leu*	pro*	HIS*	ARG*	U
	leu*	pro*	HIS*	ARG*	C
	leu	PRO*	GLN*	ARG*	A
	LEU	PRO	gln*	arg	G
A	ILE*	THR*	ASN*	SER	U
	ILE*	THR*	ASN*	SER*	C
	ile*	THR*	LYS*	arg*	A
	MET*, F-MET	THR	lys	arg	G
G	VAL*	ALA*	ASP*	GLY*	U
	VAL	ALA*	ASP*	GLY*	C
	VAL*	ALA*	GLU*	GLY*	A
	VAL	ALA	glu	GLY	G

4 x 4 x 4 = 64 "*mRNA codons*": 61 coding & 3 "*termination*"

5

Alternative Genetic Codes



6

Alternative Genetic Codes (II)

7

"Universal" DNA "*transliteration*" table

	T	C	A	G	
T	F	S	Y	C	T
	F	S	Y	C	C
	L	S	*	*	A
	L	S	*	W	G
C	L	P	H	R	T
	L	P	H	R	C
	L	P	Q	R	A
	L	P	Q	R	G
A	I	T	N	S	T
	I	T	N	S	C
	I	T	K	R	A
	M	T	K	R	G
G	V	A	D	G	T
	V	A	D	G	C
	V	A	E	G	A
	V	A	E	G	G

64 "*triplets*": 61 coding & 3 "*stops*" (*)

8

(I) *Enumeration* of "5&1" solutions

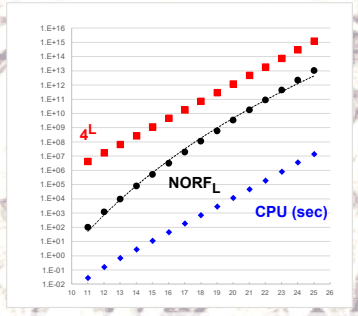
L	NORF(L)	4^L	ratio
10	0	1,048,576	-
11	96	4,194,304	43,691
12	1,152	16,777,216	14,564
13	9,216	67,108,864	7,282
14	76,320	268,435,456	3,517
15	511,104	1,073,741,824	2,101
16	3,122,688	4,294,967,296	1,375
17	19,286,112	17,179,869,184	891
18	108,498,048	68,719,476,736	633
19	588,598,272	274,877,906,944	467
20	3,204,880,608	1,099,511,627,776	343
21	16,526,184,576	4,398,046,511,104	266
22	83,667,290,112	17,592,186,044,416	210
23	424,013,102,496	70,368,744,177,664	166
24	2,072,865,313,536	281,474,976,710,656	136
25	10,026,040,699,392	1,125,899,906,842,620	112

No* "*5&1*" solutions for L ≤ 10 bp : *QED
96 for **L=11 bp**
 # Open Readings Frames, length **L** ($NORF_L$) increases ***exponentially***
Solutions are rare wrt 4^L:
Random search for exemplars ***inefficient***

9

Enumerated "5&1" $NORF_L$ solutions

versus **total sequences** (4^L) & required **CPU time**

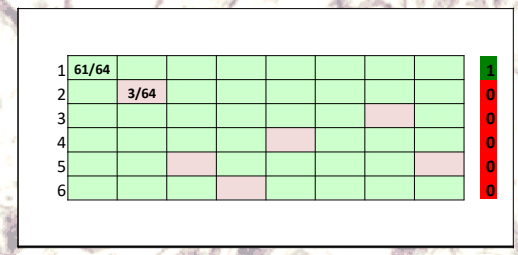


Exponential enumeration succumbs to **storage** and (or) **CPU** limitations **L~25**
How does $NORF_L$ behave wrt 4^L if **L >> 25**?

10

(II) Triplet Approximation: *Theory*

Genetic Code comprises **61** coding & **3** stop triplets:
Disregard: triplet composition, reading frame overlap, 5'➔3', etc.
p(**coding**) = **61/64** p(**stop**) = **3/64**



11

(II) Triplet Approximation: *Computation*

Consider **dsDNA** length **L bp** with **T = L / 3** triplets
P (random triplet coding) = **C = 61 / 64**

For any *string* of **T** random triplets:

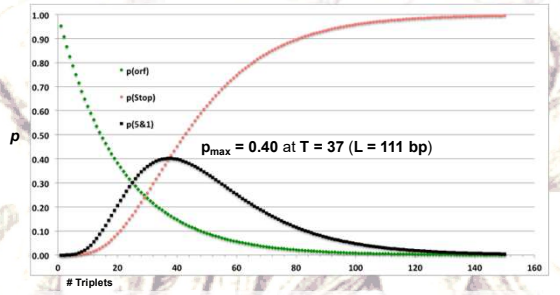
1) Prob. of an **ORF**: **p(ORF) = C^T**
2) Prob. of **≥ 1 stops**: **p(Stop) = 1 – C^T**
3) Prob. **1** Open **&** **5** Closed **= $(C^T)(1 – C^T)^5$**

Then: Joint Prob. for *any of six* RFs **p(5&1) = (6)$(C^T)(1 – C^T)^5$**

12

(II) Triplet approximation: Components

p(5&1) = (6)(pORF)(pStop)^5

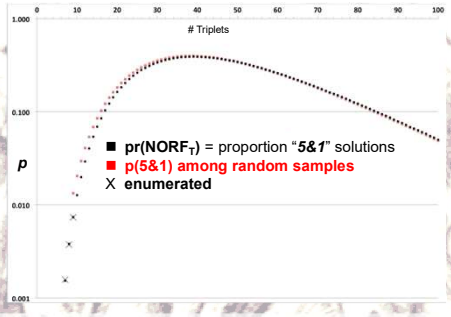


That is: **0.1 Kbp dsDNA** has **optimal (40%) prob.** of comprising a ***single*** **ORF**
5&1 solutions limited by **"multiple open"**, ***not*** **"all closed"** frames
Optimum size for **Natural Selection** to acquire / maintain / extend function ?

13

(III) Triplet simulation:

Take 10^6 random samples of **dsDNA** length **T**



Approximation provides good upper bound for **Monte Carlo Simulation**
[Simulation is **CPU-intensive**]

14

(IV) Alternative Genetic Codes

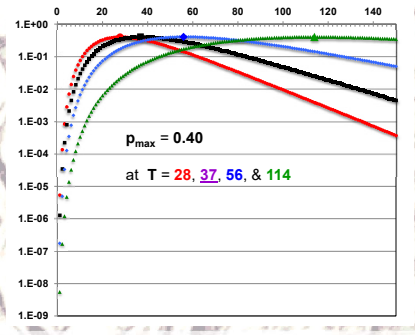
	T	C	A	G	
T	F	S	Y	C	T
	F	S	Y	C	C
	L	S	Y	W	A
	L	S	*	W	G
C	L	P	H	R	T
	L	P	H	R	C
	L	P	Q	R	A
	L	P	Q	R	G
A	I	T	N	S	T
	I	T	N	S	C
	I	T	K	R	A
	M	T	K	R	G
G	V	A	D	G	T
	V	A	D	G	C
	V	A	E	G	A
	V	A	E	G	G

"*Universal*" code has 3 stops: what if there were 4, 2, or 1 ?

15

p(5&1) for Genetic Codes

with **S** = **1**, **2**, **3**, or **4** stop triplets



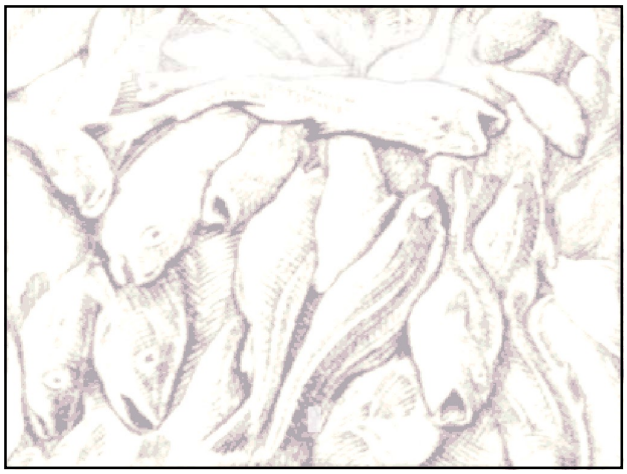
Why three stops? Does **S = 3** optimize something ?

16

Summary:

- ***Carr's Conjecture*** answered: *no* **ORF** exemplars **< 11 bp**
- **Enumeration** of exemplars is **exponential** & **CPU limited**
- **Approximation** of triplets shows
  - **$NORF_L$** function has unexpected shape & implications
    - Initially limited by improbability of simultaneous closed frames
    - Ultimately limited by improbability of **ORFs**
- **Monte Carlo sampling** of **dsDNA** space confirms approximation
- **Alternative Genetic Codes** with **n** = **1**, **2**, **4** stops *vs* **3**
  - **Random ORFs** are not rare: **P_{max} = 0.40**
  - Optimum size varies: **111** bp *vs* 84 ~ 352 bp
- Applications of **Computational Science** to **Biological Science**
- *"There's an app for that:"* ***RandomORF***
  - Carr, Wareham, & Craig (2014) *CBE Life Sci Educ* 13,373
  - [http://www.ucs.mun.ca/~donald/orf/biocomp/]

17



18

(V) Screen capture of *RandomORF*

(Carr, Wareham, & Craig 2014 *CBE Life Sci Educ* 13, 68)

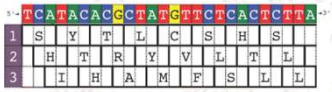
19

Bioinformatics of DNA

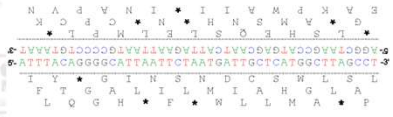
(I) DNA ***sense*** **strand** has ***same information content*** as **mRNA:**

- **5'- G T A A T C C T C - 3' DNA sense strand**
- **5'- G U A A U C C U C - 3' mRNA**
- **N - val - ile - leu - C protein**

(II) Each **DNA** strand has *three* **Reading Frames:**

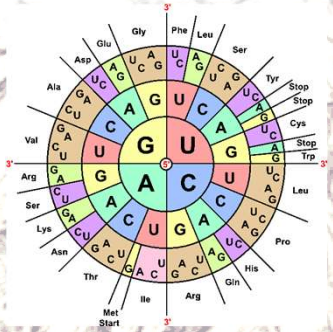


(III) Five of six 3-letter **reading frames** are "***closed***" by "***stop***" triplets.
Sixth is an ***"Open" Reading Frame***" (***ORF***) without stops that specifies a **polypeptide**



20

Alternative arrangements Genetic Code



64 "*codons*": 61 coding & 3 "*stops*" (*)

21

"*Cracking*" *the* Genetic Code

	U	C	A	G	
U					U C A G
C					U C A G
A					U C A G
G					U C A G

How far do you get with
AAA CCC GGG UUU ACAC AGAG AUAU CGCG CUCU GUGU

22

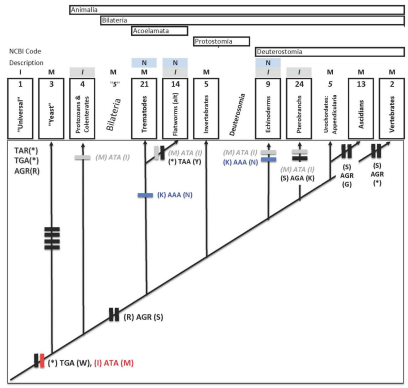
Alternative mtDNA Genetic Codes in Animalia: 1, 2, or 4 Stops

Code	Decription	1 2 3	T T T T C C C C A A A A G T A A G T T T T T A G G G T A G A T C A G A A A G G
1	Universal		F * * * L L L L L I K R R G
3	Yeast		. . . W T T T T M . . . .
4	Protozoan & Coelenterate		. . . W . . . .
5	Invertebrates		. . . W . . . . M . S S .
21	Trematodes		. . . W . . . . M N S S .
9	Echinoderms		. . . W . . . . . N S S .
14	Ascidians (alt)		. Y . W . . . . . N S S .
13	Ascidians		. . . W . . . . M . G G .
24	Pterobranchs		. . . W . . . . . . S K .
2	Vertebrates		. . . W . . . . M . * * .

NCBI DNA Alphabet Codes: 10 variable sites, 4 at Stops

23

Phylogeny of mtDNA Genetic Codes in Animals



24