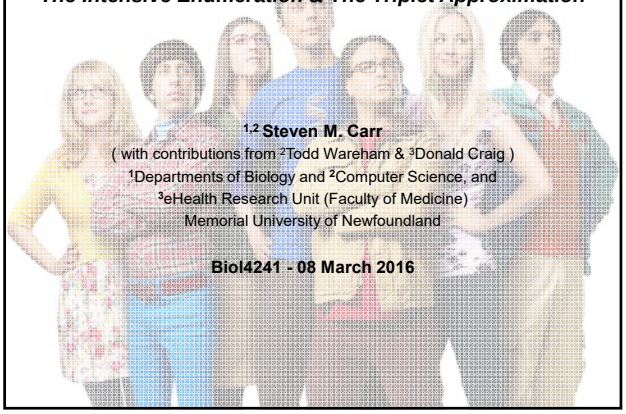
The Big ORF Theory:

The Intensive Enumeration & The Triplet Approximation



[1,2] Steven M. Carr

(with contributions from [2]Todd Wareham & [3]Donald Craig)

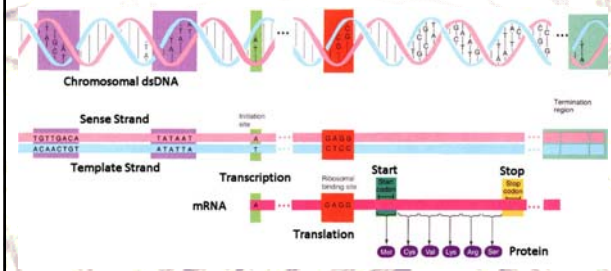
[1]Departments of Biology and [2]Computer Science, and

[3]eHealth Research Unit (Faculty of Medicine)

Memorial University of Newfoundland

Biol4241 - 08 March 2016

The Central Dogma:
DNA ➔ RNA ➔ Protein



DNA self-replicating, **double-stranded helix**
mRNA transcribed from ***Template*** **strand**
Protein translated from **mRNA codons:**
Protein reads *directly* from triplets in complementary *Sense* strand

The Genetic Code, 1965
(Nobel Prize, 1968)

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

64 RNA *codons*: 61 coding & 3 "*Termination*"

A 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*" (*)

Bioinformatics of DNA

In a **protein-coding** gene region,
five of six three-letter **reading frames** are "***closed***" by "***stop***" triplets,
one is "***Open Reading Frame***" (***ORF***) without "*stops*" that specifies a **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 and bioinformatic education:
I want an app for that !!!

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

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

Pseudo-code for
Skeleton, Complete, & *Generate* functions

```
function GenerateSkeleton(S, ORFNum, rfn)
  if rfn += 7 then
    return CompleteSequence(ORFNum, S)
  elif rfn == ORFNum
    return GenerateSkeleton(S, ORFNum, rfn + 1)
  else
    res = Null
    for (pos, stopCodon) pair in a randomization of the
        list of all possible such pairs for reading frame rfn
        of S do
      St = place stop codon stopCodon at position pos relative
          to reading frame rfn of S
      if that stop placement is possible then
        res = GenerateSkeleton(St, ORFNum, rfn + 1)
        if res is not equal to Null then
          exit for-loop
    return res

function CompleteSequence(ORFNum, S)
  if S has no more unfilled bases then
    return S
  else
    res = Null
    pos = position of unfilled base in S
    for base in randomization of list ["A", "G", "C", "T"] do
      St = S
      St[pos] = base
      if the number of stops in open reading frame of St equals 0 then
        res = CompleteSequence(ORFNum, St)
        if res is not equal to Null then
          exit for-loop
    return res

function generateRandomORF(seqLen)
  let S be a sequence of length seqLen of unfilled bases
  ORFNum = random selection from reading-frame numbers 1 .. 6
  return GenerateSkeleton(S, ORFNum, 1)
```

(I) Enumeration of "5&1" solutions

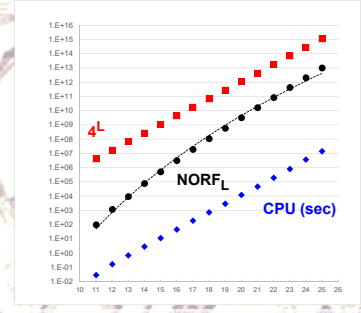
L	NORF(L)	4^L	ratio
10	0	1,048,576	-
11	96	4,194,304	43,691
12	1,152	16,777,216	14,564
13	9,216	67,108,864	7,282
14	76,320	268,435,456	3,517
15	511,104	1,073,741,824	2,101
16	3,122,688	4,294,967,296	1,375
17	19,286,112	17,179,869,184	891
18	108,498,048	68,719,476,736	633
19	588,598,272	274,877,906,944	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

There are no "***5&1***" solutions for **L ≤ 10 bp** : ***QED***
96 for **L=11 bp** ; **NORF$_L$** increases exponentially

Solutions are rare wrt **4^L**:
Random search for exemplars is **inefficient**

Enumerated "5&1" solutions length L (NORF$_L$)

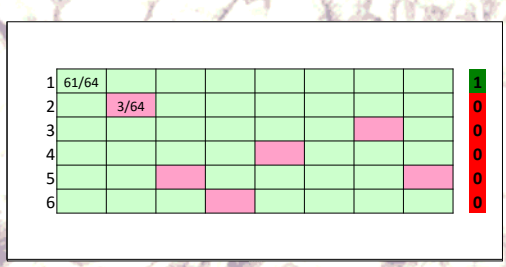
versus total sequences (**4^L**) & required **CPU time**



Enumeration succumbs to **storage** and (or) **CPU** limitations @ **L~25**
How does **NORF$_L$** behave wrt **4^L** if **L >> 25**?

(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**



(II) Triplet Approximation: Computation

Consider a **dsDNA** length **L bp** with **T = L / 3** triplets
Genetic Code comprises **61** coding & **3** stop triplets:
Disregard triplet composition, reading frame overlap, etc.
Prob that random triplet is coding is **C = 61/64**

For any set (string) of **T** random triplets:

1) Prob of an **ORF**: **p(ORF) = C^T**

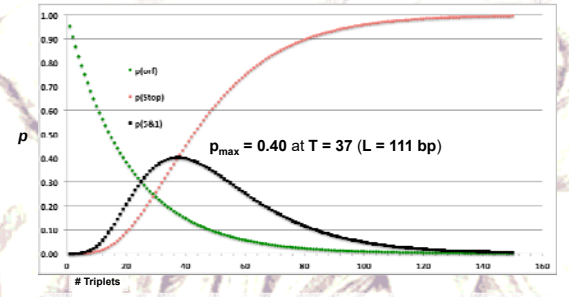
2) Prob of **≥ 1 stops**: **p(Stop) = 1 – C^T**

3) Prob RF1 Open ***&*** RFs2-6 closed **= $(C^T)(1 – C^T)^5$**

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

(II.i) Triplet approximation:

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

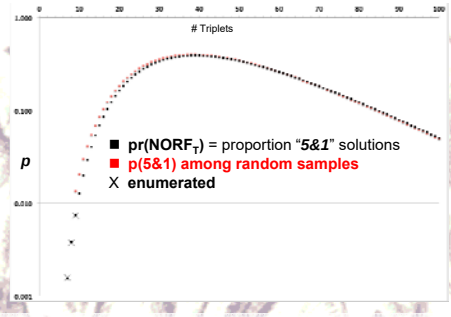


p_{max} = 0.40 at **T = 37** (**L = 111 bp**)

0.1 Kbp dsDNA has **optimal** (**40%**) **chance** of including **single ORF**
Solutions limited *first* by **"multiple open"**, *later* by **"all closed"** frames
Optimum size for **Natural Selection** to acquire / maintain / extend function?

(II.ii) Triplet simulation:

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



■ **pr(NORF$_T$)** = proportion "***5&1***" solutions
■ **p(5&1) among random samples**
X **enumerated**

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

(III) 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

There are 3 stops: what if there were 4, 2, or 1 ?

Evolution of animal mtDNA Genetic Codes: Triplet re-assignments

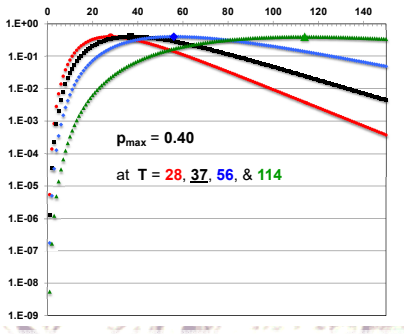
Triplet Position												
1	T	T	T	C	C	C	C	A	A	A	A	
2	A	A	G	T	T	T	T	T	A	G	G	
3	A	G	A	T	C	A	G	A	A	A	G	
Code	Phylogenetic occurrence											
1	**Universal**	*	*	*	L	L	L	L	I	K	R	R
4	**Ophisthokonta** [Fungi + Animalia]	.	.	W	.	.	.	.	.	.	.	.
3	**Fungi** (inc Yeast)	.	.	W	T	T	T	T	M	.	.	.
5	**Protostomia**, Urochordata I [Appendicularia]	.	.	W	.	.	.	.	M	.	S	S
9	**Acoelomata** [Platyhelminthes], **Echinodermata**, Hemichordata I [Enteropneusta]	.	.	W	.	.	.	.	.	N	S	S
14	*Trichobilharzia, Radolphus*	Y	.	W	.	.	.	.	.	N	S	S
24	Hemichordata II [Pterobranchia]	.	.	W	.	.	.	.	.	.	S	K
13	Urochordata II [Ascidea + Thalicea]	.	.	W	.	.	.	.	M	.	G	G
2	**Chordata**	.	.	W	.	.	.	.	M	.	*	*

NCBI GenBank codes, October 2014

Which (Code) came first, the **Acoelomates** or the **Protostomes** ?

p(5&1) for genetic codes

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



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

Summary:

- ***Carr's Conjecture*** is answered: *no* **ORF** exemplars **< 11 bp**
- **Enumeration of *"5&1"* DNA exemplars** is **exponential & CPU limited**
- **Triplet approximation** shows
  - **$NORF_L$** function has unexpected shape & implications
    - Initially limited by improbability of simultaneous closed frames
    - Ultimately limited by improbability of **ORF**
  - Alternative Genetic Codes with **n** = 3 *vs* 1, 2, or 4 stop codons
    - **Random ORFs** are not rare: **P_{max} = 0.40**
    - Optimum size varies 84 ~ (111) ~ 352 bp
  - Confirmed by random sampling of **dsDNA** space
- ***"There's an app for that:" RandomORF***
  - Carr, Craig, & Wareham (2014) *CBE Life Sci Educ* 13,68
  - [http://www.ucs.mun.ca/~donald/orf/biocomp/]

Acknowledgements

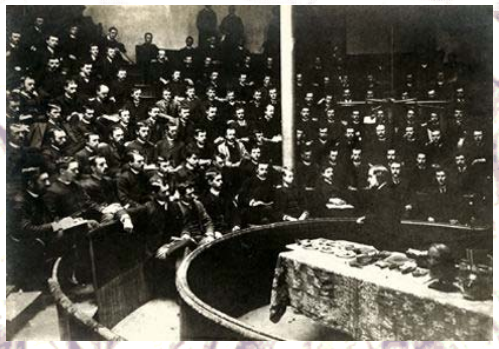
Todd Wareham, Dept of Computer Science &

Donald Craig, eHealth Research Unit (Faculty of Medicine)

Memorial University of Newfoundland

NSERC Discovery Grants to Drs Carr & Wareham

Acknowledgements



Students in Biol2250 – Principles of Genetics