Basic mathematical genomics

Nick Landolfi and Dan O'Neill
Stanford University

Revision: January 2022

DNA

- deoxyribonucleic acid (DNA) is a molecule which we can represent as a string
- a *nucleotide* (or *base pair*) is one of adenine, thymine, cytosine, and guanine
  - we abbreviate the word "nucleotide" with "nt"
  - we represent each of the four nucleotides with letters A, T, C, and G
  - ribonucleic acid (RNA) has the nucleotide uracil (U) instead of thymine (T)
- a *nucleotide string* is a sequence in the set $\{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$; e.g, ATCGATCATC
  - in the "double helix" structure of DNA, A binds with T and C binds with G, forming "cross-bars"
  - we call A the *nucleotide complement* of T, and vice versa; same for C and G
  - as a result, we can represent the double helix DNA as a single nucleotide string

Proteins

- a protein is a molecule which we can represent as a string
- an *amino acid* (also *residue*) is one of

name	symbol	name	symbol	name	symbol	name	symbol
alanine	A	arginine	R	asparagine	N	aspartate	D
cysteine	C	glutamine	Q	glutamate	E	glycine	G
histidine	H	isoleucine	I	leucine	L	lysine	K
methionine	M	phenylalanine	F	proline	P	serine	S
threonine	T	tryptophan	W	tyrosine	Y	valine	V

- an *amino acid string* is a sequence in {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}
  - we denote this set by $\mathcal{A}$, a mnemonic for "amino"
  - different amino acid strings correspond to different proteins
  - as a result, we can represent a protein as a single amino acid string

Codons

- nucleotides have semantic meaning in *non-overlapping* sequences of three
- a *nucleotide codon* (or *trinucleotide sequence*) is a length 3 nucleotide string; e.g., ATC
  - codons encode an element of $\mathcal{A}$ (an amino acid) or a "stop" (which we denote by $\diamond$)
  - we partition the set $\{\mathsf{A}, \mathsf{T}, \mathsf{C}, \mathsf{G}\}^3$ of $4^3 = 64$ codons into 61 *amino codons* and 3 *stop codons*
- a nucleotide string is *codon-aligned* if its length is a multiple of three
  - a codon-aligned nucleotide string can be interpreted as a sequence of codons
  - we know the *codon decoding function* $f : \{\mathsf{A}, \mathsf{C}, \mathsf{T}, \mathsf{G}\}^3 \to \mathcal{A} \cup \{\diamond\}$
    - for example, $f(\mathsf{GCT}) = \mathsf{A}$ where the r.h.s. is the symbol for the amino alanine
  - f is not injective since two distinct codons may map to the same amino (or to $\diamond$)
    - we call two codons with the same image under f *synonyms*
    - for example, CAU and CAC are synonyms for histidine; i.e., $f(\mathsf{CAU}) = f(\mathsf{CAC}) = \mathsf{H}$

Codon Table

- it is easier to tabulate f^{-1} since its codomain is smaller than its domain

symbol	codons; i.e., $f^{-1}(\text{symbol})$	symbol	codons
A	GCT, GCC, GCA, GCG	I	ATT, ATC, ATA
R	CGT, CGC, CGA, CGG, AGA, AGG	L	CTT, CTC, CTA, CTG, TTA, TTG
N	AAT, AAC	K	AAA, AAG
D	GAT, GAC	M	ATG
C	TGT, TGC	F	TTT, TTC
Q	CAA, CAG	P	CCT, CCC, CCA, CCG
E	GAA, GAG	S	TCT, TCC, TCA, TCG, AGT, AGC
G	GGT, GGC, GGA, GGG	T	ACT, ACC, ACA, ACG
H	CAT, CAC	W	TGG
$\diamond$	TAA, TGA, TAG	Y	TAT, TAC
		V	GTT, GTC, GTA, GTG

- the domain of f is $\{\text{A}, \text{T}, \text{C}, \text{G}\}^3$ and the codomain of f is $\mathcal{A} \cup \{\diamond\}$
- $f^{-1}(x)$ is the set of domain elements of f (in this case, codons) which map to $x \in \mathcal{A} \cup \{\diamond\}$

Nucleotide senses

- naturally, we can extend f to codon-aligned nucleotide strings by defining $s = \bar{f}(x)$ by

$$s_i = f(\underbrace{x_{3(i-1)+1}x_{3(i-1)+2}x_{3(i-1)+3}}_{\text{codon } i \text{ of } x})$$

  - we call s the sense of x; for example, the sense of ATTCTTAAA is

$$\bar{f}(\underbrace{\text{ATT}}_{\text{I}}\,\underbrace{\text{CTT}}_{\text{L}}\,\underbrace{\text{AAA}}_{\text{K}}) = \text{ILK}$$

- since f is not one-to-one, neither is $\bar{f}$
  - x and y are *sense-equivalent* if they have the same sense; i.e., $\bar{f}(x) = \bar{f}(y)$
  - roughly speaking, x and y are sense-equivalent if they “spell out the same thing”
  - e.g., CGTCGC and CGACGG are sense-equivalent because $\bar{f}(\underbrace{\text{CGT}}_{\text{R}}\,\underbrace{\text{CGC}}_{\text{R}}) = \bar{f}(\underbrace{\text{CGA}}_{\text{R}}\,\underbrace{\text{CGG}}_{\text{R}}) = \text{RR}$
    - in this case, because CGT, CGC, CGA, CGG are synonyms for arginine (R)

Nucleotide substitutions

- a *(nucleotide) substitution* (or *point mutation*) to a length m nucleotide string is a pair (j, b)
  - the *index* j is in $\{1, \ldots, m\}$ and the *replacement* nucleotide b is in $\{\text{A}, \text{T}, \text{C}, \text{G}\}$
  - the *(j, b)-mutation* of x is the nucleotide string y defined by $y_j = b$ and $y_i = x_i$ for all $i \neq j$
    - i.e., y is the same as x except at index j, where it has nucleotide b
    - e.g., the (3,A)-mutation of $\text{CG}\underline{\text{T}}$ is $\text{CG}\underline{\text{A}}$ (we swapped T in position 3 with A)
- we classify substitutions on codon-aligned nucleotide sequences by their effect on the sense
  - a substitution is *synonymous (silent)* if it does not change the sense
    - e.g. (3,C) on $\text{CG}\underline{\text{T}}$ with result $\text{CG}\underline{\text{C}}$, since $f(\text{CG}\underline{\text{T}}) = f(\text{CG}\underline{\text{C}}) = \text{R}$
  - a substitution is *nonsynonymous* if it changes the sense
    - a substitution is *missense* if an amino codon became a different amino codon
      - the *missense variants* of a protein are all proteins which differ with it by one amino in position
    - a substitution is *nonsense (readstop)* if an amino codon became a stop codon
    - a substitution is *nonstop (readthrough)* if a stop codon became amino codon