Deciphering the genetic code of overlapping genes

S. Lèbre et O. Gascuel

Université de Strasbourg, ICube – UMR 7357
Institut de Biologie Computationnelle, LIRMM, CNRS & Université de Montpellier
Overlapping genes

1977: a single DNA sequence may code for several overlapping genes

Genetic code: DNA bases \{a,c,g,t\} amino \Rightarrow 1 amino acid

One sequence
- different reading frame
- same or opposite strain

5 frames
Overlapping genes

• First (70s)
  • non-viral species
  • multiple functions: regulation, translational coupling, genome imprinting...

• Recently
  • Number of overlapping genes could be greater than expected
  • Especially in the virus world
    HIV: 3 overlapping genes (env, tat, rev)
    HTLV-1 (Human T-cell leukemia virus): HBZ gene

• Favored therapeutic targets
  ➔ *Highly conserved* DNA sequences are subject to strong evolutionary constraints
  ➔ *Prevent* the *rapid adaptation* of viruses
    and fast appearance of *resistance mutations*. 
Degrees of freedom?

• Mathematical results in the 1980s
  • Sander and Schultz (1979)
  • Siegel and Fitch ((1980)
  • Smith and Waterman (1980) : conditional information
    ➔ Frame dependent
    ➔ Frame \( f = -2 \) : « very rare in nature »

• But: HBZ (HTLV), ASP (HIV) overlap in \( f = -2 \)

• Still questions...

• Looking for explicit constraints:
  1) For 2 overlapping “proteins” \((\text{symmetric view})\)
  2) When 1 protein is known \((\text{asymmetric view})\)

➔ amino-acid (n-peptide) composition
➔ detecting selection pressure
➔ searching for overlapping genes
Opposite frames $F^-$

Let $R_f$ be the relation such that $xR_f y$ whenever sequence $y$ overlap the reference sequence $x$ with frame $f$

$$f \in F^- = \{-2, -1, 0\} : \quad xR_f y \iff yR_f x$$
Same sense frames $F^+$

• Let $R_f$ be the relation such that $xR_f y$ whenever sequence $y$ overlap the reference sequence $x$ with frame $f$

$$f \in F^- = \{-2, -1, 0\}: \quad xR_f y \iff yR_f x$$

$$F^+ = \{+1, +2\}: \quad xR_{+1} y \iff yR_{+2} x$$
Amino acid constraints

• Frame $f = -0$
  (opposite strand, without shift)
  $\rightarrow$ 5 constraints

• From the genetic code (without stop in the 2 reading frames)
  1) Reference frame ‘aac’ (N) $\rightarrow$ ‘gtt’ (v) in overlapping frame
  2) ‘aat’ (N) $\rightarrow$ ‘att’ (i)
  3) ‘gtc’, ‘gta’, ‘gtg’ (v) $\rightarrow$ ‘gac’ (D), ‘tac’ (Y), ‘cac’ (H)
  4) ‘atc’, ‘ata’ (i) $\rightarrow$ ‘gat’ (D), ‘tat’ (Y)
  5) … $\leftarrow$ …

• But for the other overlapping frames?

Partial codon overlap  $\rightarrow$ Dependency
Linear Algebraic approach

- **Quadons**

  In all frames \( f \neq -0 \), 4 DNA bases describe 2 overlapping codons (amino acids in both reading frame)

- **Vector** \( \mathbf{Q} \) of size \((4^4 - \#\text{stops})\) gives the number of occurrences of **quadons** or 4-letter words in the sequence (except Stops)

- **Vector** \( \mathbf{N} \) of size 40 gives the number of occurrences of 20 amino acids (without a stop) in both frames (reference and overlap),

\[
\mathbf{N} = (\ [A]_1, \ [C]_1, \ [D]_1, \ ...[Y]_1, \ [A]_2, \ [C]_2, \ [D]_2, \ ...,[Y]_2 ).
\]
Linear Algebraic approach

\[
\begin{bmatrix}
A_1 \\
\vdots \\
T_1 \\
\vdots \\
Y_1 \\
A_2 \\
\vdots \\
V_2 \\
\vdots \\
Y_2
\end{bmatrix}
= 
\begin{bmatrix}
\cdots & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & 0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\
\end{bmatrix}
\begin{bmatrix}
\text{aaaa} \\
\text{aaac} \\
\text{...} \\
\text{aact} \\
\text{...} \\
\text{atat} \\
\text{...} \\
\text{...} \\
\text{tttt}
\end{bmatrix}
\]

\[N_{[40]} = M_f^{\left[40 \times (4^4 - 24)\right]} Q_{[4^4 - 24]}\]

- # of AA (Ref & overlap)
- # of quadons (- 24 Stop)
\( \text{Remaining degrees of freedom} \)

- \( M_f \) has not full rank:
  \[ \sum_{i=1}^{20} L_i = \sum_{i=1}^{20} L_{20+i} \quad (\text{Trivial constraint}) \]

- This may be the only constraint

\textit{Frame } f = -1 : \textit{only one (trivial) linear constraint} between reference/overlapping protein amino acid composition

- For all other frame shifts, additional constraints do exist

\[ \text{Equality constraints correspond to the set of linear combinations of the lines of matrix } M_f \]

\[ \text{Number of equality constraints} = 2 \times 20 - \text{Rank}(M_f) \]
Constraints list for all frames

Frame $f = -2 \Rightarrow 10$ constraints

(Ref)

- $A = a$
- $Y = y$
- $G = p$
- $P = g$
- $T = v$
- $V = t$
- $H + Q = c + w$
- $C + W = h + q$
- $I + M = i + m$

+the trivial constraint ($F + L + S + N + K + D + E + R$)

Symmetric

Extension to di-peptides

Frame $f = -2 \Rightarrow 125$ di-pep constraints

2x 6 null constraints (2 frames)
- $FY = 0, YY = 0, HY = 0, NY = 0, DY = 0, CY = 0$

(⇒ STOP in overlapping: $YY$ \{‘tat’, ‘tac’\}*2 overlap in frame $f=-2$ with ‘taa’ or ‘tag’\)}
Number of equality constraints

Amino acid

-2: 10
-1: 1
0: 5
+1: 2
+2: 2

2-peptide

-2: 113
-1: 1
0: 25
+1: 4
+2: 4

3-peptide

-2: 1316
-1: 1
0: 125
+1: 8
+2: 8

- Null constraint = STOP in at least 1 frame

- Tri-peptides
- Higher order (Graph traversal algorithm)
Normalized number of equality constraints

Normalized number of constraints = $\sqrt[\text{n}]{C_n}$
where $C_n$ is the # of constraints for peptides of length $n$
Average number of amino acid choice

\[
S_{n}^{f} = \left( \frac{1}{\#\text{Pep}} \sum_{c=1}^{\#\text{constraints}} |Pep_{1,c}^{f}| \right) \left( \frac{1}{Pep_{2,c}^{f}} \right)
\]

• Example
  (2-peptide constraints, \( f = -2 \))

\[
S_{n=2}^{f=-2} = \sqrt{\frac{1+1+0+0+2 \times 2 + 2 \times 2 + 6 \times 4 + ...}{20^2}}
\]

AA = AA
AG = PA
PA = AG
YY = 0
0 = YY
AH + AQ = CA + WA
CA + WA = AH + AQ
YF + YL + YS + YN + YK + YR = LY + KY + EY + RY
...

Average number of AA choice
due to sets of equality constraints
When one protein is given...

Local n-peptide constraints

Average # of AA choice due to local constraints

\[ S_n^f = \left( \frac{1}{20^n} \sum_{i=1}^{20^n} |Pep_i^f| \right)^{\frac{1}{n-1}} \]

Smith & Waterman (1980)

| Reading frame m | \( I_m(C|C) \) | \( I_m(C|C \times C) \) |
|-----------------|----------------|------------------|
| Ref             | 4.218          | —                |
| +2              | 2.144          | 1.709            |
| +1              | 2.144          | 1.729            |
| -0              | 1.532          | —                |
| -1              | 3.424          | 1.832            |
| -2              | 0.821          | 0.644            |

TABLE 2

The Average Conditional Information per Codon
 Obtained from Eqs. (8) and (9)
 When the Encoding of Each Amino Acid
 Defines a Codon Class.
a 10th protein in HIV virus?

- env : gene coding for the virus capside
- ASP may be a protein coded by a gene overlapping env with $f = -2$
- ASP
  - 189 amino acids: 103 fixed + 86 flexibles
  - Average Hydrophobicity (Kyte Doolittle)
    - Fixed AA: 0.9
    - Flexible AA:
      - Observed: 0.9
      - Expected: 0.2 ($\sigma^2 = 0.06$)

Frame $f = -2$
env aa composition
HIVb aa frequency

Highly Hydrophobic
Conclusion

• 2 points of view
  • 2 proteins: peptide equality constraints
  • 1 known protein

• Tools for studying pression selection?

 ===> Poster

*Evolutionary analyses strongly support that ASP (Anti Sense Protein) overlapping ORF is the 10th gene of HIV-1 M pandemic group*

Elodie Cassan, Anne-Muriel Chifolleau, Antoine Gross, Olivier Gascuel.