Back-translation for discovering distant protein homologies

Marta Gîrdea, Laurent Noé, Gregory Kucherov

Université Lille 1 - CNRS - INRIA (France)

WABI 2009, University of Pennsylvania, Philadelphia, September 12-13, 2009
Back-translation

Back-translation of Amino acid R (Arginine) and its set of codons graph.

Back-translation of Amino acids YSH and its back-translation graph.

Marta Gîrdea, Laurent Noé, Gregory Kucherov. Back-translation for discovering distant protein homologies 2/25
Amino acid R (Arginine) and its set of codons (graph):

\[
\begin{align*}
\text{R} & \quad \text{Back-translation} \\
\begin{align*}
C \to G \to A & \quad C \to A \\
C \to G \to C & \quad G \to C \\
C \to G \to G & \quad G \to G \\
C \to G \to T & \quad C \to G \to T \\
A \to G \to A & \quad A \to G \to A \\
A \to G \to G & \quad G \to G
\end{align*}
\end{align*}
\]
Back-translation

- Amino acid R (Arginine) and its set of codons (graph):

- Amino acids YSH and its back-translation graph:
Back-translation

previous work

1. Compute the best back-translated sequence using multiple protein alignment.

[Moreira and Maass, 2004, Giugno et al., 2004]

DNA curvature, absence of interactions & restriction sites

[Gonnet, 2005]

2. Back-translation & frameshifts

i. Half-related work: use BLASTN to predict frameshifts

[Raes and Van de Peer, 2005, Okamura, 2006, Harrison and Yu, 2007, Hahn and Lee, 2005]

→ not a "back-translation" since you also need DNA sequence amino acid substitution scores based on DNA similarities

[Leluk, 1998, Leluk, 2000]

→ was not designed for frameshifted alignment

ii. Related work: amino acid score matrices with frameshifts

[Pellegrini and Yeates, 1999]

→ does not predict frameshift inside proteins

aligning sequence graphs

[Arvestad, 1997, Arvestad, 2000]

→ alignment of translated codons with all possible frameshifts

→ time costly algorithm
Compute the best *back-translated sequence* using
- multiple protein alignment
  [Moreira and Maass, 2004, Giugno et al., 2004]
- DNA curvature, absence of interactions & restrictions sites
  [Gonnet, 2005]
Back-translation

previous work

1. Compute the best *back-translated sequence* using
   - multiple protein alignment
     [Moreira and Maass, 2004, Giugno et al., 2004]
   - DNA curvature, absence of interactions & restrictions sites
     [Gonnet, 2005]

2. Back-translation & *frameshifts*
Compute the best *back-translated sequence* using
- multiple protein alignment
  [Moreira and Maass, 2004, Giugno et al., 2004]
- DNA curvature, absence of interactions & restrictions sites
  [Gonnet, 2005]

Back-translation & *frameshifts*

i. Half-related work:
- use BLASTN to predict frameshifts [Raes and Van de Peer, 2005, Okamura, 2006, Harrison and Yu, 2007, Hahn and Lee, 2005]
  → not a “back-translation” since you also need DNA sequence
- amino acid substitution scores based on DNA similarities
  [Leluk, 1998, Leluk, 2000]
  → was not designed for frameshifted alignment
Compute the best *back-translated sequence* using
- multiple protein alignment
  [Moreira and Maass, 2004, Giugno et al., 2004]
- DNA curvature, absence of interactions & restrictions sites
  [Gonnet, 2005]

Back-translation & frameshifts
i. Half-related work:
- use BLASTN to predict frameshifts [Raes and Van de Peer, 2005, Okamura, 2006, Harrison and Yu, 2007, Hahn and Lee, 2005]
  → not a “back-translation” since you also need DNA sequence
- amino acid substitution scores based on DNA similarities
  [Leluk, 1998, Leluk, 2000]
  → was not designed for frameshifted alignment

ii. Related work:
- amino acid *score matrices with frameshifts* [Pellegrini and Yeates, 1999]
  → does not predict frameshift *inside* proteins
- aligning *sequence graphs* [Arvestad, 1997, Arvestad, 2000]
  → alignment of translated codons with all possible frameshifts
  → time costly algorithm
Back-translation alignment

find the "best" alignment of DNA sequences that encode the target proteins

Marta Gîrdea, Laurent Noé, Gregory Kucherov
find the “best” alignment of DNA sequences that encode the target proteins
find the “best” alignment of DNA sequences that encode the target proteins

AGN: \[ C \rightarrow C \rightarrow G \rightarrow G \rightarrow A \rightarrow A \rightarrow C \]

QET: \[ C \rightarrow A \rightarrow G \rightarrow A \rightarrow A \rightarrow C \rightarrow G \rightarrow T \]

Marta Gîrde, Laurent Noé, Gregory Kucherov
find the “best” alignment of DNA sequences that encode the target proteins

**AGN:** $C \rightarrow C \rightarrow C \rightarrow G \rightarrow G \rightarrow C \rightarrow A \rightarrow A \rightarrow C$

**QET:** $C \rightarrow A \rightarrow G \rightarrow G \rightarrow A \rightarrow A \rightarrow C \rightarrow C \rightarrow T$

**Alignment**

$C \rightarrow C \rightarrow A \rightarrow G \rightarrow G \rightarrow A \rightarrow A \rightarrow A \rightarrow C$

$C \rightarrow A \rightarrow G \rightarrow G \rightarrow A \rightarrow A \rightarrow A \rightarrow C \rightarrow A \rightarrow C \rightarrow G \rightarrow T$
Usage scenarios

- Hidden homologies (virus overlapped genes)
- Frameshifts & incorrect translations (programmed frameshifts, biological or “human” errors).
- ...
Could be done by classic coding DNA alignment, but:
- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model
Could be done by classic coding DNA alignment, but:

- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model

...GCCTGTCTCATCATGGAAGGCGCTGAATTTACGGAAG...
Could be done by classic coding DNA alignment, but:

- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model
Could be done by classic coding DNA alignment, but:
- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model
Back-translation alignment

Could be done by classic coding DNA alignment, but:

- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model
Could be done by classic coding DNA alignment, but:
- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model
Could be done by classic coding DNA alignment, but:

- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model
Back-translation alignment algorithm

\[ M[i, j, (\alpha_i, \beta_j)] = \max \begin{cases} 0 (a) & M[i-1, j-1, (\alpha_i-1, \beta_j-1)] + \text{score}(\alpha_i, \beta_j), \alpha_i-1 \in \text{pred}_G A(\alpha_i); \\ (b) & M[i, j-1, (\alpha_i, \beta_j-1)] + \text{singleGapPenalty}, \beta_j-1 \in \text{pred}_G B(\beta_j); \\ (c) & M[i-1, j, (\alpha_i-1, \beta_j)] + \text{singleGapPenalty}, \alpha_i-1 \in \text{pred}_G A(\alpha_i); \\ (d) & M[i, j-3, (\alpha_i, \beta_j-3)] + \text{tripleGapPenalty}, j \geq 3 \\ (e) & M[i-3, j, (\alpha_i-3, \beta_j)] + \text{tripleGapPenalty}, i \geq 3 \end{cases} \]

where \( G_A \) and \( G_B \) are the back-translated graphs being aligned.
\( \alpha_i \) (respectively \( \beta_j \)) is a labelled node at position \( i \) (resp. \( j \)) of \( G_A \) (resp. \( G_B \)).
\( \text{pred}_G(n) \) is the set of nodes that precede \( n \) on the back-translated graph \( G \).

In practice, singleGapPenalty and tripleGapPenalty are affine gap functions.
Extended Smith-Waterman algorithm on two back-translation graphs

\[
M[i,j, (\alpha_i, \beta_j)] = \max \begin{cases} 
0 (a) \\
M[i-1, j-1, (\alpha_i-1, \beta_j-1)] + \text{score}(\alpha_i, \beta_j), \alpha_i-1 \in \text{pred}_G A(\alpha_i) \); \\
M[i, j-1, (\alpha_i, \beta_j-1)] + \text{singleGapPenalty}, \beta_j-1 \in \text{pred}_G B(\beta_j) \); \\
M[i-1, j, (\alpha_i-1, \beta_j)] + \text{singleGapPenalty}, \alpha_i-1 \in \text{pred}_G A(\alpha_i) \); \\
M[i-3, j, (\alpha_i-3, \beta_j-3)] + \text{tripleGapPenalty}, j \geq 3 \); \\
M[i-3, j, (\alpha_i-3, \beta_j)] + \text{tripleGapPenalty}, i \geq 3 \). 
\end{cases}
\]
Extended *Smith-Waterman* algorithm on two back-translation graphs

\[
M[i, j,(\alpha_i, \beta_j)] =
\begin{cases}
0 & \\
M[i - 1, j - 1,(\alpha_{i-1}, \beta_{i-1})] + \text{score}(\alpha_i, \beta_j), & \alpha_{i-1} \in \text{pred}_{G_A}(\alpha_i); \\
M[i, j - 1,(\alpha_i, \beta_{j-1})] + \text{singleGapPenalty}, & \beta_{j-1} \in \text{pred}_{G_B}(\beta_j); \\
M[i - 1, j,(\alpha_{i-1}, \beta_j)] + \text{singleGapPenalty}, & \alpha_{i-1} \in \text{pred}_{G_A}(\alpha_i); \\
M[i, j - 3,(\alpha_i, \beta_{j-3})] + \text{tripleGapPenalty}, & j \geq 3 \\
M[i - 3, j,(\alpha_{i-3}, \beta_j)] + \text{tripleGapPenalty}, & i \geq 3
\end{cases}
\]

where

- \( G_A \) and \( G_B \) are the back-translated graphs being aligned
- \( \alpha_i \) (respectively \( \beta_j \)) is a labelled node at position \( i \) (resp. \( j \)) of \( G_A \) (resp. \( G_B \))
- \( \text{pred}_G(n) \) is the set of nodes that precede \( n \) on the back-translated graph \( G \).
Back-translation alignment algorithm

Extended Smith-Waterman algorithm on two back-translation graphs

\[ M[i, j, (\alpha_i, \beta_j)] = \]
\[
\begin{cases} 
0, & (a) \\
M[i - 1, j - 1, (\alpha_{i-1}, \beta_{j-1})] + \text{score}(\alpha_i, \beta_j), & (b) \\
\max \left\{ M[i, j - 1, (\alpha_i, \beta_{j-1})] + \text{singleGapPenalty}, (c) \\
M[i - 1, j, (\alpha_{i-1}, \beta_j)] + \text{singleGapPenalty}, (d) \\
M[i, j - 3, (\alpha_i, \beta_{j-3})] + \text{tripleGapPenalty}, \beta_{j-1} \in \text{pred}_{G_B}(\beta_j); (e) \\
M[i - 3, j, (\alpha_{i-3}, \beta_j)] + \text{tripleGapPenalty}, \alpha_{i-1} \in \text{pred}_{G_A}(\alpha_i); (f) \\
\right. 
\end{cases}
\]

where

- \( G_A \) and \( G_B \) are the back-translated graphs being aligned
- \( \alpha_i \) (respectively \( \beta_j \)) is a labelled node at position \( i \) (resp. \( j \)) of \( G_A \) (resp. \( G_B \))
- \( \text{pred}_G(n) \) is the set of nodes that precede \( n \) on the back-translated graph \( G \).

In practice, \text{singleGapPenalty} and \text{tripleGapPenalty} are affine gap functions
Back-translation alignment algorithm

Extended *Smith-Waterman* algorithm on two back-translation graphs
Back-translation alignment algorithm

Extended *Smith-Waterman* algorithm on two back-translation graphs

Each DP matrix cell is composed of several entries

\[ M[i, j] \]

\((\alpha_i, \beta_j)\)
Extended *Smith-Waterman* algorithm on two back-translation graphs
Back-translation alignment algorithm

Extended *Smith-Waterman* algorithm on two back-translation graphs

Simple,
Back-translation alignment algorithm

Extended *Smith-Waterman* algorithm on two back-translation graphs

Simple, but it does not work ...
Back-translation alignment algorithm

Extended *Smith-Waterman* algorithm on two back-translation graphs

Simple, but it does not work ...

**Reason:** the scoring system
Reasons:
Scoring system

Reasons:

- at the nucleic level, at least 1/4 of the matches are non significant.
Scoring system

Reasons:

- at the nucleic level, at least 1/4 of the matches are non significant.
- in the context of back-translated sequences:
  - some matches can be easily obtained (3\textsuperscript{rd} codon position),
  - other are much more difficult to get.
Scoring system

Reasons:

- at the nucleic level, at least 1/4 of the matches are non significant.
- in the context of back-translated sequences:
  - some matches can be easily obtained (3\textsuperscript{rd} codon position),
  - other are much more difficult to get.

→ matching context plays an important role.
Our scoring system depends on:

1. the amino acids being aligned,
2. the nucleic positions in the corresponding codons,
3. the nucleic bases at these positions.

Moreover, it distinguishes the actual codons being aligned (no ambiguity).
Scoring system

Our scoring system depends on:

1. the **amino acids** being aligned,
Our scoring system depends on:

1. the **amino acids** being aligned,
2. the **nucleic positions** in the corresponding codons,
Scoring system

Our scoring system depends on:

1. the amino acids being aligned,
2. the nucleic positions in the corresponding codons,
3. the nucleic bases at these positions.
Our scoring system depends on:

1. the **amino acids** being aligned,

2. the **nucleic positions** in the corresponding codons,

3. the **nucleic bases at these positions**.

Moreover, it distinguishes the *actual codons* being aligned (no ambiguity)
Scoring system

“Evolutionary” point of view

codons
shared coding sequence translated on two reading frames

duplication

independent divergence

pair of symbols with “common origins”
Our scoring matrices are computed as *log odd ratio* of such evolutionary scenario based on substitution models:

1. Goldman model [Kosiol et al., 2007] → mechanical substitution model, no AA constraints
2. “codon-PAM” model [Schneider et al., 2005] → empirical substitution model on vertebrates, thus with AA replacement constraints
Our scoring matrices are computed as \textit{log odd ratio} of such evolutionary scenario based on substitution models:

1. Goldman model [Kosiol et al., 2007] → mechanical substitution model, no AA constraints
2. “codon-PAM” model [Schneider et al., 2005] → empirical substitution model on vertebrates, thus with AA replacement constraints

Each time, we compute 6 scoring matrices according to the codon position (0,1,2) on both sequences being aligned.
Our scoring matrices are computed as log odd ratio of such evolutionary scenario based on substitution models:

1. Goldman model [Kosiol et al., 2007] → mechanical substitution model, no AA constraints
2. “codon-PAM” model [Schneider et al., 2005] → empirical substitution model on vertebrates, thus with AA replacement constraints

Each time, we compute 6 scoring matrices according to the codon position (0,1,2) on both sequences being aligned.

E-value is computed with the Karlin’s λ and K parameters (Island Method [Altschul and et al, 2001]).
### Scoring system

Substitution scores between nucleotides located at position 1 (vertical) and 0 (horizontal) respectively of the codons in the aligned backtranslated proteins.

| 1<->0 | A|I | A|K | A|M | A|N | A|R | A|S | A|T | C|H | C|L | C|P | C|Q | C|R | G|A | G|D | G|E | G|G |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A|D | 1.23 | -0.26 | 1.23 | -0.17 | -0.12 | 0.07 | 1.28 | -1.38 | -0.64 | -1.07 | -1.29 | -0.73 | -0.37 | -1.69 | -1.64 | -1.35 |
| A|E | -0.41 | 1.08 | -0.34 | 1.06 | 1.27 | 1.17 | -0.04 | -0.65 | -1.54 | -1.44 | -0.63 | 0.58 | -1.08 | -0.94 | -1.11 | -0.66 |
| A|H | 1.39 | -0.40 | 1.48 | -0.29 | -0.23 | 0.08 | 1.55 | -1.13 | -0.34 | -0.69 | -1.04 | -0.74 | 0.02 | -1.24 | -1.12 | -0.85 |
| A|K | -1.16 | 1.04 | -1.04 | 0.99 | 1.16 | 1.05 | -0.52 | -0.68 | -1.75 | -1.51 | -0.61 | 0.47 | -0.82 | -0.17 | -0.17 | 0.07 |
| A|N | 1.13 | -0.82 | 1.18 | -0.66 | -0.70 | -0.34 | 1.24 | -1.10 | -0.42 | -0.62 | -1.04 | -1.01 | 0.04 | -1.53 | -1.39 | -1.15 |
| A|Q | -0.63 | 1.08 | -0.61 | 0.97 | 1.46 | 1.34 | -0.12 | -0.41 | -1.27 | -1.09 | -0.39 | 0.79 | -0.79 | -0.62 | -0.67 | -0.17 |
| A|Y | 1.44 | -1.33 | 1.53 | -0.94 | -1.30 | -0.40 | 1.59 | -1.73 | -0.34 | -0.94 | -1.57 | -1.76 | -0.19 | -2.21 | -2.00 | -1.55 |
| C|A | -0.74 | -0.80 | -0.56 | -0.97 | -0.02 | -0.97 | -0.86 | 1.00 | 0.87 | 1.17 | 0.90 | 0.63 | -0.68 | -1.44 | -1.12 | -1.25 |
| C|P | -0.74 | -0.63 | -0.50 | -0.88 | 0.19 | -0.92 | -0.93 | 1.24 | 1.01 | 1.21 | 1.13 | 0.86 | -0.99 | -1.73 | -1.25 | -1.58 |
| C|S | -0.62 | -0.82 | -0.44 | -1.06 | 0.06 | -1.09 | -0.81 | 1.08 | 0.98 | 1.23 | 1.02 | 0.76 | -0.62 | -1.72 | -1.27 | -1.48 |
| C|T | -0.65 | -0.74 | -0.49 | -0.94 | 0.06 | -0.95 | -0.74 | 1.01 | 0.77 | 1.06 | 0.94 | 0.70 | -0.59 | -1.57 | -1.19 | -1.39 |
| G|C | 0.44 | -1.69 | -0.02 | -1.39 | -1.51 | -0.87 | 0.11 | -1.32 | -0.43 | -0.44 | -1.16 | -1.48 | 1.65 | -1.44 | -1.12 | -0.58 |
| G|G | -0.24 | -0.99 | -0.54 | -0.61 | -0.84 | -0.42 | -0.30 | -0.85 | -0.88 | -0.67 | -0.63 | -0.97 | 0.98 | 0.93 | 0.97 | 1.04 |
| G|R | -0.10 | -0.23 | -0.34 | -0.01 | -0.07 | 0.23 | -0.04 | -0.98 | -1.05 | -0.84 | -0.58 | -0.59 | 1.12 | 1.23 | 1.35 | 1.48 |
| G|S | 0.31 | -1.13 | 0.10 | -1.00 | -0.83 | -0.69 | 0.21 | -0.66 | -0.24 | -0.10 | -0.62 | -0.74 | 1.17 | -0.98 | -0.80 | -0.51 |
| G|W | -1.94 | -1.64 | -1.71 | -0.69 | -0.70 | 0.22 | -1.16 | -1.62 | -2.34 | -1.77 | -1.58 | -1.23 | -0.34 | -0.71 | -0.78 | 2.56 |
| T|F | -0.47 | -2.28 | -0.24 | -1.74 | -2.11 | -1.02 | -0.10 | -1.45 | 0.29 | -0.56 | -1.55 | -1.90 | -0.46 | -2.60 | -2.49 | -1.82 |
| T|I | -1.25 | -2.11 | -0.94 | -1.70 | -1.79 | -1.14 | -0.71 | -0.81 | 0.08 | -0.56 | -1.16 | -1.41 | -0.72 | -2.39 | -2.32 | -1.94 |
| T|L | -1.51 | -1.96 | -1.25 | -1.64 | -1.43 | -1.00 | -1.03 | -0.80 | -0.40 | -1.01 | -1.25 | -1.11 | -1.10 | -2.45 | -2.42 | -1.68 |
| T|M | -1.88 | -1.74 | -1.70 | -1.62 | -1.04 | -0.95 | -1.58 | -0.67 | -0.92 | -1.15 | -0.91 | -0.67 | -1.57 | -2.28 | -2.20 | -1.49 |
| T|V | -1.34 | -1.80 | -1.06 | -1.55 | -1.23 | -1.00 | -0.90 | -0.48 | -0.02 | -0.46 | -0.76 | -0.79 | -0.88 | -2.22 | -2.13 | -1.64 |
| g|R | -0.83 | 0.03 | -0.95 | 0.21 | 0.09 | 0.29 | -0.58 | -0.84 | -1.53 | -1.25 | -0.48 | -0.46 | 0.30 | 1.38 | 1.48 | 1.55 |
### Scoring system

Substitution scores between nucleotides located at position 1 (vertical) and 0 (horizontal) respectively of the codons in the aligned backtranslated proteins

| 1<->0 | A | I | A | K | A | M | A | N | A | R | A | S | A | T | C | H | C | L | C | P | C | Q | C | R | G | A | G | D | G | E | G | G |
|-------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | D | 1.23 | -0.26 | 1.23 | -0.17 | -0.12 | 0.07 | 1.28 | -1.38 | -0.64 | -1.07 | -1.29 | -0.73 | -0.37 | -1.69 | -1.64 | -1.35 |
| A | E | -0.41 | 1.08 | -0.34 | 1.06 | 1.27 | 1.17 | -0.04 | -0.65 | -1.54 | -1.44 | -0.63 | 0.58 | -1.08 | -0.94 | -1.11 | -0.66 |
| A | H | 1.39 | -0.40 | 1.48 | -0.29 | -0.23 | 0.08 | 1.55 | -1.13 | -0.34 | -0.69 | -1.04 | -0.74 | 0.02 | -1.24 | -1.12 | -0.85 |
| A | K | -1.16 | 1.04 | -1.04 | 0.99 | 1.16 | 1.05 | -0.52 | -0.68 | -1.75 | -1.51 | -0.61 | 0.47 | -0.82 | -0.17 | -0.17 | 0.07 |
| A | N | 1.13 | -0.82 | 1.18 | -0.66 | -0.70 | -0.34 | 1.24 | -1.10 | -0.42 | -0.62 | -1.04 | -1.01 | 0.04 | -1.53 | -1.39 | -1.15 |
| A | Q | -0.63 | 1.08 | -0.51 | 0.97 | 1.46 | 1.34 | -0.12 | -0.41 | -1.27 | -1.09 | -0.39 | 0.79 | -0.79 | -0.62 | -0.67 | -0.17 |
| A | Y | 1.44 | -1.33 | 1.53 | -0.94 | -1.30 | -0.40 | 1.59 | -1.73 | -0.34 | -0.94 | -1.57 | -1.76 | -0.19 | -2.21 | -2.00 | -1.55 |
| C | A | -0.74 | -0.80 | -0.56 | -0.97 | -0.02 | -0.97 | -0.86 | 1.00 | 0.87 | 1.17 | 0.90 | 0.63 | -0.68 | -1.44 | -1.12 | -1.25 |
| C | P | -0.74 | -0.63 | -0.50 | -0.88 | 0.19 | -0.92 | -0.93 | 1.24 | 1.01 | 1.21 | 1.13 | 0.86 | -0.99 | -1.73 | -1.25 | -1.58 |
| C | S | -0.62 | -0.82 | -0.44 | -1.06 | 0.06 | -1.09 | -0.81 | 1.08 | 0.98 | 1.23 | 1.02 | 0.76 | -0.62 | -1.72 | -1.27 | -1.48 |
| C | T | -0.65 | -0.74 | -0.49 | -0.94 | 0.06 | -0.95 | -0.74 | 1.01 | 0.77 | 1.06 | 0.94 | 0.70 | -0.59 | -1.57 | -1.19 | -1.39 |
| G | C | 0.44 | -1.69 | -0.02 | -1.39 | -1.51 | -0.87 | 0.11 | -1.32 | -0.43 | -0.44 | -1.16 | -1.48 | 1.65 | -1.44 | -1.12 | -0.58 |
| G | G | -0.24 | -0.99 | -0.54 | -0.61 | -0.84 | -0.42 | -0.30 | -0.85 | -0.88 | -0.67 | -0.63 | -0.97 | 0.98 | 0.93 | 0.97 | 1.04 |
| G | R | -0.10 | -0.23 | -0.34 | -0.01 | -0.07 | 0.23 | -0.04 | -0.98 | -1.05 | -0.84 | -0.58 | -0.59 | 1.12 | 1.23 | 1.35 | 1.48 |
| G | S | 0.31 | -1.13 | 0.10 | -1.00 | -0.83 | -0.69 | 0.21 | -0.66 | -0.24 | -0.10 | -0.62 | -0.74 | 1.17 | -0.98 | -0.80 | -0.51 |
| G | W | -1.94 | -1.64 | -1.71 | -0.69 | -0.70 | 0.22 | -1.16 | -1.62 | -2.34 | -1.77 | -1.58 | -1.23 | -0.34 | -0.71 | -0.78 | 2.56 |
| T | F | -0.47 | -2.28 | -0.24 | -1.74 | -2.11 | -1.02 | -0.10 | -1.45 | 0.29 | -0.56 | -1.55 | -1.90 | -0.46 | -2.60 | -2.49 | -1.82 |
| T | I | -1.25 | -2.11 | -0.94 | -1.70 | -1.79 | -1.14 | -0.71 | -0.81 | 0.08 | -0.56 | -1.16 | -1.41 | -0.72 | -2.39 | -2.32 | -1.94 |
| T | L | -1.51 | -1.96 | -1.25 | -1.64 | -1.43 | -1.00 | -1.03 | -0.80 | -0.40 | -1.01 | -1.25 | -1.11 | -1.10 | -2.45 | -2.42 | -1.68 |
| T | M | -1.88 | -1.74 | -1.70 | -1.62 | -1.04 | -0.95 | -1.58 | -0.67 | -0.92 | -1.15 | -0.91 | -0.67 | -1.57 | -2.28 | -2.20 | -1.49 |
| T | V | -1.34 | -1.80 | -1.06 | -1.55 | -1.23 | -1.00 | -0.90 | -0.48 | -0.02 | -0.46 | -0.76 | -0.79 | -0.88 | -2.22 | -2.13 | -1.64 |
| g | R | -0.83 | 0.03 | -0.95 | 0.21 | 0.09 | 0.29 | -0.58 | -0.84 | -1.53 | -1.25 | -0.48 | -0.46 | 0.30 | 1.38 | 1.48 | 1.55 |

Aligning the second Guanine of a codon “W” (Tryptophan) : G W
against the first Guanine of a codon “G” (Glycine) : G G
is an “exceptional” event.
Scoring system
Substitution scores between nucleotides located at position 1 (vertical) and 0 (horizontal) respectively of the codons in the aligned backtranslated proteins

| 1<->0 | A|D | A|E | A|H | A|K | A|N | A|Q | A|Y | C|A | C|P | C|S | C|T | G|C | G|G | G|R | G|S | G|W | T|F | T|I | T|L | T|M | T|V | g|R |
|-------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|       |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| A|D   | 1.23| -0.26| 1.23| -0.17| -0.12| 0.07| 1.28|-1.38|-0.64|-1.07|-1.29|-0.73|-0.37|-1.69|-1.64|-1.35|
| A|E   | -0.41| 1.08| -0.34| 1.06| 1.27| 1.17|-0.04|-0.65|-1.54|-1.44|-0.63| 0.58|-1.08|-0.94|-1.11|-0.66|
| A|H   | 1.39| -0.40| 1.48| -0.29|-0.23| 0.08| 1.55|-1.13|-0.34|-0.69|-1.04|-0.74| 0.02|-1.24|-1.12|-0.85|
| A|K   | -1.16| 1.04|-1.04| 0.99| 1.16| 1.05|-0.52|-0.68|-1.75|-1.51|-0.61| 0.47|-0.82|-0.17| 0.17| 0.07|
| A|N   | 1.13| -0.82| 1.18| -0.66|-0.70|-0.34| 1.24|-1.10|-0.42|-0.62|-1.04|-1.01| 0.04|-1.53|-1.39|-1.15|
| A|Q   | -0.63| 1.08|-0.51| 0.97| 1.46| 1.34|-0.12|-0.41|-1.27|-1.09|-0.39| 0.79|-0.79|-0.62|-0.67|-0.17|
| A|Y   | 1.44| -1.33| 1.53| -0.94|-1.30|-0.40| 1.59|-1.73|-0.34|-0.94|-1.57|-1.76|-0.19|-2.21|-2.00|-1.55|
| C|A   | -0.74| -0.80|-0.56|-0.97|-0.02|-0.97|-0.86| 1.00| 0.87| 1.17| 0.90| 0.63| -0.68|-1.44|-1.12|-1.25|
| C|P   | -0.74| -0.63|-0.50|-0.88| 0.19|-0.92|-0.93| 1.24| 1.01| 1.21| 1.13| 0.86| -0.99|-1.73|-1.25|-1.58|
| C|S   | -0.62| -0.82|-0.44|-1.06| 0.06|-1.09|-0.81| 1.08| 0.98| 1.23| 1.02| 0.76| -0.62|-1.72|-1.27|-1.48|
| C|T   | -0.65| -0.74|-0.49|-0.94| 0.06|-0.95|-0.74| 1.01| 0.77| 1.06| 0.94| 0.70| -0.59|-1.57|-1.19|-1.39|
| G|C   | 0.44| -1.69|-0.02|-1.39|-1.51|-0.87| 0.11|-1.32|-0.43|-0.44|-1.16|-1.48| 1.65|-1.44|-1.12|-0.58|
| G|G   | -0.24| -0.99|-0.54|-0.61|-0.84|-0.42|-0.30|-0.85|-0.88|-0.67|-0.63|-0.97| 0.98| 0.93| 0.97| 1.04|
| G|R   | -0.10| -0.23|-0.34|-0.01|-0.07| 0.23|-0.04|-0.98|-1.05|-0.84|-0.58|-0.59| 1.12| 1.23| 1.35| 1.48|
| G|S   | 0.31| -1.13| 0.10|-1.00|-0.83|-0.69| 0.21|-0.66|-0.24|-0.10|-0.62|-0.74| 1.17|-0.98|-0.80|-0.51|
| G|W   | -1.94| -1.64|-1.71|-0.69|-0.70| 0.22|-1.16|-1.62|-2.34|-1.77|-1.58|-1.23| -0.34|-0.71|-0.78| 2.56|
| T|F   | -0.47| -2.28|-0.24|-1.74|-2.11|-1.02|-0.10|-1.45| 0.29|-0.56|-1.55|-1.90| -0.46|-2.60|-2.49|-1.82|
| T|I   | -1.25| -2.11|-0.94|-1.70|-1.79|-1.14|-0.71|-0.81| 0.08|-0.56|-1.16|-1.41| -0.72|-2.39|-2.32|-1.94|
| T|L   | -1.51| -1.96|-1.25|-1.64|-1.43|-1.00|-1.03|-0.80|-0.40|-1.01|-1.25|-1.11|-1.10|-2.45|-2.42|-1.68|
| T|M   | -1.88| -1.74|-1.70|-1.62|-1.04|-0.95|-1.58|-0.67|-0.92|-1.15|-0.91|-0.67|-1.57|-2.28|-2.20|-1.49|
| T|V   | -1.34| -1.80|-1.06|-1.55|-1.23|-1.00|-0.90|-0.48|-0.02|-0.46|-0.76|-0.79|-0.88|-2.22|-2.13|-1.64|
| g|R   | -0.83| 0.03| -0.95| 0.21| 0.09| 0.29|-0.58|-0.84|-1.53|-1.25|-0.48|-0.46| 0.30| 1.38| 1.48| 1.55|

Aligning the second Guanine of a codon “W” (Tryptophan): G W
against the first Guanine of a codon “G” (Glycine): G G
is an “exceptional” event.

Tryptophan codons: TGG, Glycine codons: GGN
Advanced snakes venom neurotoxins

Malayan krait (*Bungarus Candidus*) & Monocled cobra (*Naja kaouthia*)

Diversification of venom toxins is studied in [Fry et al., 2008]
Advanced snakes venom neurotoxins

Malayan krait (Bungarus Candidus)

### GCAGTATGTGTATCATTATTAGGAGCAGCAAATATACCACCACATCCATTC
### AATTTAATAAATTTTATGAAGATGATAAGATATACAATA

### GCAGTATGTATCATTATTAGGAGCAGCAAATATACCACCACATCCACTC
### AATTTAATAAATTTTATGGAGATGATAAGATATACAATA

### GCATGTGAAAAAACATGGGGAGAATATGTGGATTATGGATGTTATTGTGGA
### GTGGGAGGATCAGGAAGACCAATAGATGCATTAGATAGA

### GCATGTGAAAAAACATGGGGAGAATATGCGGATTATGGATGTTATTGTGGA
### GCGGGAGGATCAGGAAGACCAATAGATGCATTAGATAGA

### GCATGTGAAAAAACATGGGGAGAATATGCTGTACATGATAATTGTTATGGAGATGCAGAAAAAAAACATAAAAA
### TGTAATCCAAAAATGCAATCATATTCATATAAATTAACA

### TCTGAATACATCGAGCGGCACAAGAATATTGACACCGCGAGATATTGCC
### SEYIEKTYICGVHELDSYSAIMTARFC

### TCTGAATACATCGAGCGGCACAAGAATATTGACACCGCGAGATTTTGCC
### SEYIEKTYICGVHELDSYSAIMTARFC
Enterobacteriaceae transposases

*Yersinia pestis*

Most probably a programmed translational frameshifting: observed in transposases of related species as in *E. coli* [Licznar et al., 2006]
Unsure frameshifts, both inside two exons: if confirmed, does not modify any important domain of the protein

Strong tips: absence of STOP codons after the first frameshift in two reading-frames + strong mRNA conservation (see next slide)
|    | PDGFA_HUMAN     | BAA00987.1     | Homo Sapiens | Rattus norvegicus |
|----|----------------|---------------|--------------|-------------------|
| 1  | MRTLACLLLLLGCGYLAHVLAEEAEIPREVI  | MRTLWACLLLLLGCGYLAH | 50           |                   |
|    | EAI   |                            |              |                   |
| 50 | PDGFA_HUMAN     | BAA00987.1     | Homo Sapiens | Rattus norvegicus |
| 51 | DSVGSEDSLDTSLRAHG | DSVGAEDALETNLRA  | 100          |                   |
|    | DPLKRPIRRKRSIEEAAP | HGHSVKHVEPKRVP |              |                   |
| 100| PDGFA_HUMAN     | BAA00987.1     | Homo Sapiens | Rattus norvegicus |
| 101| VIYEIPRSQVDPTSANFLIWPC | VIYIPEG | 150          |                   |
|    | QVEVKRCTGCCNTSSVK | AQVEVKRCTGCC |              |                   |
| 150| PDGFA_HUMAN     | BAA00987.1     | Homo Sapiens | Rattus norvegicus |
| 151| KVAKVEYVRKKPKLKEV | KVAKVEYVRKKPKLKE | 193          |                   |
|    | QVRLEEHLECA   | QVRLEEHLECA      |              |                   |
| 193| PDGFA_HUMAN     | BAA00987.1     | Homo Sapiens | Rattus norvegicus |
|    | YREEDT         | YREEDT          |              |                   |

Unsure frameshifts, both inside two exons: if confirmed, does not modify any important domain of the protein. Strong tips: absence of STOP codons after the first frameshift in two reading-frames + strong mRNA conservation (see next slide).
Mammals Platelet-derived growth factors

Homo Sapiens & Rattus norvegicus

Back-translation for discovering distant protein homologies
Mammals Platelet-derived growth factors

*Homo Sapiens & Rattus norvegicus*

Unsure frameshifts, both inside two exons: if confirmed, does not modify any important domain of the protein
Mammals Platelet-derived growth factors

Homo Sapiens & Rattus norvegicus

Unsure frameshifts, both inside two exons: if confirmed, does not modify any important domain of the protein

Strong tips: absence of STOP codons after the first frameshift in two reading-frames + strong mRNA conservation (see next slide)
Platelet-derived growth factors

**Homo Sapiens & Rattus norvegicus**

**Homo Sapiens & Danio rerio**

Strong conservation (of the protein mainly due the mRNA constraints) on a distant species Zebra fish (Danio rerio) just after the 1st frameshift.

| Homo Sapiens (PDGFA_HUMAN) | 50 | Homo Sapiens (PDGFA_HUMAN) | 50 |
|----------------------------|----|-----------------------------|----|
| MRTLACLLLLGCGYLAHVLAEEAILPREIERSQIHSIRDLQRLLEI | 100 | MRTLACLLLLGCGYLAHVLAEEAILPREIERSQIHSIRDLQRLLEI | 100 |
| BAA00987.1 1 | 50 | BAA00987.1 1 | 50 |
| MRTACLLLLGCGYLAEEAILPREIERSQIHSIRDLQRLLEI | 100 | MRTACLLLLGCGYLAEEAILPREIERSQIHSIRDLQRLLEI | 100 |
| PDGFA_DANIORE 1 | 100 | PDGFA_DANIORE 1 | 100 |
| MRTALIHFLVCCSLSAAAEAPIPREIERLSNIEHSISLQIREM | 100 | MRTALIHFLVCCSLSAAAEAPIPREIERLSNIEHSISLQIREM | 100 |
| --- | --- | --- | --- |
| PDGFA_HUMAN 51 | 100 | PDGFA_HUMAN 51 | 100 |
| DSVDLAGSLSLVHGHATKHCNLPLSRIRKRSIEEAVPAVCKTRT | 150 | DSVDLAGSLSLVHGHATKHCNLPLSRIRKRSIEEAVPAVCKTRT | 150 |
| BAA00987.1 51 | 100 | BAA00987.1 51 | 100 |
| DSVDLAGSLSLVHGHATKHCNLPLSRIRKRSIEEAVPAVCKTRT | 150 | DSVDLAGSLSLVHGHATKHCNLPLSRIRKRSIEEAVPAVCKTRT | 150 |
| PDGFA_DANIORE 51 | 150 | PDGFA_DANIORE 51 | 150 |
| DFLNEVLEDVQGHKHEHLYDSRLK-LHSSKRSIEEAVPAVCKTRT | 150 | DFLNEVLEDVQGHKHEHLYDSRLK-LHSSKRSIEEAVPAVCKTRT | 150 |
| --- | --- | --- | --- |
| PDGFA_HUMAN 101 | 150 | PDGFA_HUMAN 101 | 150 |
| VIEIPRSSQDPTSANFILWPPCVEVRKRTGCGCNCNTTSSSVCQPSVHRHSV | 193 | VIEIPRSSQDPTSANFILWPPCVEVRKRTGCGCNCNTTSSSVCQPSVHRHSV | 193 |
| BAA00987.1 101 | 150 | BAA00987.1 101 | 150 |
| VIEIPRSSQDPTSANFILWPPCVEVRKRTGCGCNCNTTSSSVCQPSVHRHSV | 193 | VIEIPRSSQDPTSANFILWPPCVEVRKRTGCGCNCNTTSSSVCQPSVHRHSV | 193 |
| PDGFA_DANIORE 101 | 193 | PDGFA_DANIORE 101 | 193 |
| VIIEIPRSSQDPTSANFILWPPCVEVRKRTGCGCNCNTTSSSVCQPSVHRHSV | 193 | VIIEIPRSSQDPTSANFILWPPCVEVRKRTGCGCNCNTTSSSVCQPSVHRHSV | 193 |
| --- | --- | --- | --- |
| PDGFA_HUMAN 151 | 193 | PDGFA_HUMAN 151 | 193 |
| KVAKEVRKPKLKEVQVRLEHLACATTSNLNPYREEDT | 193 | KVAKEVRKPKLKEVQVRLEHLACATTSNLNPYREEDT | 193 |
| BAA00987.1 151 | 193 | BAA00987.1 151 | 193 |
| KVAKEVRKPKLKEVQVRLEHLACATTSNLNPYREEDT | 193 | KVAKEVRKPKLKEVQVRLEHLACATTSNLNPYREEDT | 193 |
| PDGFA_DANIORE 151 | 193 | PDGFA_DANIORE 151 | 193 |
| KVAKEVRKPKLKEVQVRLEHLACATTSNLNPYREEDT | 193 | KVAKEVRKPKLKEVQVRLEHLACATTSNLNPYREEDT | 193 |
Proposed tool
Available at http://bioinfo.lifl.fr/path/
Proposed tool
Available at http://bioinfo.lifl.fr/path/
A new method to discover hidden protein homologies:

1. algorithm that detects frameshifts on distant proteins
2. associated substitution matrices and significance parameters
Conclusion & Future Work
Available at http://bioinfo.lifl.fr/path/

A new method to discover hidden protein homologies:

1. algorithm that detects frameshifts on distant proteins
2. associated substitution matrices and significance parameters

Future work:

1. low complexity filtering (both Protein and Codon ...)
2. multiple alignment (to quickly confirm a frameshift ...)
3. seeding techniques for back-translation graphs (speed up ...)
4. large scale studies of frameshift events (takes lot of CPU-time ...)

Marta Gîrdea, Laurent Noé, Gregory Kucherov
Thank you for your attention

http://bioinfo.lifl.fr/path/
Altschul, S. and et al (2001). The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research*, 29(2):351–361.

Arvestad, L. (1997). Aligning coding DNA in the presence of frame-shift errors. *Proceedings of the 8th Annual CPM Symposium*, 1264:180–190.

Arvestad, L. (2000). *Algorithms for biological sequence alignment*. PhD thesis, Royal Institute of Technology, Stockholm, Numerical Analysis and Computer Science.

Fry, B. G., Scheib, H., van der Weerd, L., Young, B., McNaughtan, J., Ryan Ramjan, S. F., Vidal, N., Poelmann, R. E., and Norman, J. A. (2008). Evolution of an arsenal: Structural and functional diversification of the venom system in the advanced snakes (caenophidia). *Molecular and Cellular Proteomics*, 7:215–246.

Giugno, R., Pulvirenti, A., Ragusa, M., Facciola, L., Patelmo, L., Di Pietro, V., Di Pietro, C., Purrello, M., and Ferro, A. (2004). Locally sensitive backtranslation based on multiple sequence alignment. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, (CIBCB)*, pages 231–237.
Gonnet, G. H. (2005). 
Back translation (protein to DNA) in an optimal way. 
Technical Report 505, Informatik, ETH, Zurich.

Hahn, Y. and Lee, B. (2005). 
Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. 
*Bioinformatics*, 21(Suppl 1):i186–i194.

Harrison, P. and Yu, Z. (2007). 
Frame disruptions in human mRNA transcripts, and their relationship with splicing and protein structures. 
*BMC Genomics*, 8:371.

Kosiol, C., Holmes, I., and Goldman, N. (2007). 
An Empirical Codon Model for Protein Sequence Evolution. 
*Molecular Biology and Evolution*, 24(7):1464.

Leluk, J. (1998). 
A new algorithm for analysis of the homology in protein primary structure. 
*Computers and Chemistry*, 22(1):123–131.

Leluk, J. (2000). 
A non-statistical approach to protein mutational variability. 
*BioSystems*, 56(2-3):83–93.
Licznar, P., Bertrand, C., Canal, I., Prère, M.-F., and Fayet, O. (2006). Genetic variability of the frameshift region in IS911 transposable elements from escherichia coli clinical isolates. *FEMS Microbiology Letters*, 218(2):231–237.

Moreira, A. and Maass, A. (2004). TIP: protein backtranslation aided by genetic algorithms. *Bioinformatics*, 20(13):2148.

Okamura, K. et al. (2006). Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics*, 88(6):690–697.

Pellegrini, M. and Yeates, T. (1999). Searching for Frameshift Evolutionary Relationships Between Protein Sequence Families. *Proteins*, 37:278–283.

Raes, J. and Van de Peer, Y. (2005). Functional divergence of proteins through frameshift mutations. *Trends in Genetics*, 21(8):428–431.

Schneider, A., Cannarozzi, G., and Gonnet, G. (2005). Empirical codon substitution matrix. *BMC bioinformatics*, 6(1):134.