

# Back-translation for discovering distant protein homologies

Marta Gîrdea, **Laurent Noé**, Gregory Kucherov

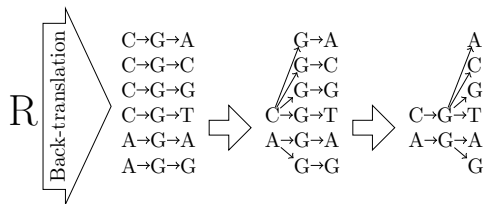
Université Lille 1 - CNRS - INRIA (France)

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

# Back-translation

# Back-translation

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





# Back-translation

previous work

# 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]

# 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*

- 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*
  - 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

- 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*
  - 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

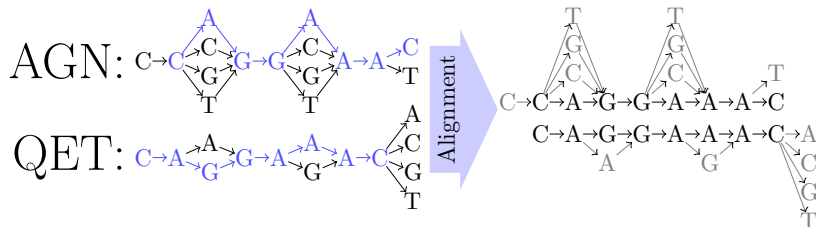
# Back-translation alignment

find the “best” alignment of DNA sequences that encode the target proteins



# Back-translation alignment

find the “best” alignment of DNA sequences that encode the target proteins



# Back-translation alignment

## Usage scenarios

- Hidden homologies (virus overlapped genes)
- Frameshifts & incorrect translations (programmed frameshifts, biological or “human” errors).
- ...

# 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

# 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

· · · GCCTGTCTCATCATGGAAGGCGCTGAATTTACGGAAG · · ·

# 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

· · · GCCTGTCTCATCATGGAAGGCGCTGAATTTACGGAAG · · ·

A C L I M E G A E F T E

# 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

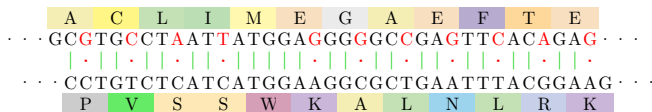
	A	C	L	I	M	E	G	A	E	F	T	E																										
· · ·	G	C	C	T	G	T	C	T	C	A	T	C	A	T	G	G	A	A	G	G	C	G	C	T	G	A	A	T	T	T	A	C	G	G	A	A	G	· · ·
	P	V	S	S	W	K	A	L	N	L	R	K																										



# Back-translation alignment

Could be done by classic coding DNA alignment, but:

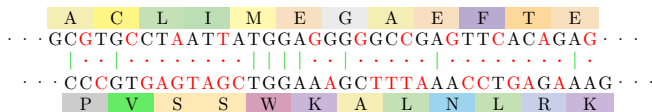
- coding DNA evolves faster than Protein
- synonymous mutations are “free” in our model



# 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



# Back-translation alignment algorithm

# 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

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

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

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

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

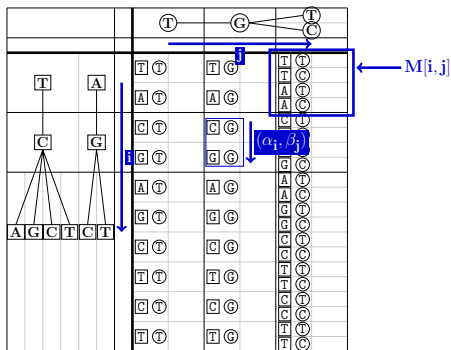
# Back-translation alignment algorithm

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



# 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

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:**

## Reasons:

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

## 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 ( $3^{rd}$  codon position),
  - other are much more difficult to get.

## 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<sup>rd</sup> codon position),
  - other are much more difficult to get.

→ matching context plays an important role.

# Scoring system

Our scoring system depends on:

# Scoring system

Our scoring system depends on:

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

# Scoring system

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**.

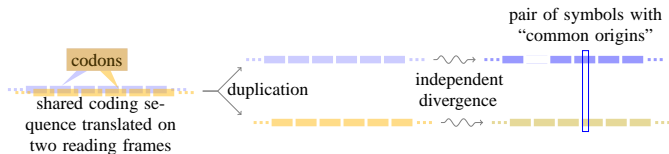
Our scoring system depends on:

- 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



# Scoring system

## Substitution matrices

- 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

# Scoring system

## Substitution matrices

- 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.

# Scoring system

## Substitution matrices

- 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  $\lambda$  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	-0.51	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	0.05	-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	0.05	-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 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	0.05	-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*)



```
A V C V S L L G A A N I P P H P F N L I N F M K M I R Y T I  
GCAGTATGTGTATCATTATTAGGAGCAGCAAAATATACCACCACATCCATTCAATTTAATAAAATTTTATGAAGATGATAAGATATACAATA  
GCAGTATGTGTATCATTATTAGGAGCAGCAAAATATACCACCACATCCACTCAATTTAATAAAATTTTATGGAGATGATAAGATATACAATA  
A V C V S L L G A A N I P P H P L N L I N F M E M I R Y T I
```

```
F C E K T W G E Y V D Y G C Y C G V G G S G R P I D A L D R  
CCATGTGAAAAACATGGGGAGAAATATGGGATATGGATGTTATTTGGAGTGGGAGGATCAGGAAGACCAATAGATGCATTAGATAGA  
CCATGTGAAAAACATGGGGAGAAATATGGGGATATGGATGTTATTTGGAGCGGAGGATCAGGAAGACCAATAGATGCATTAGATAGA  
F C E K T W G E Y A D Y G C Y C G A G G S G R P I D A L D R
```

```
C C Y V H D N C Y G D A E K K H K C N P K M Q S Y S Y K L T  
TGTTGTTATGTACATGATAATGTTATGGAGATGCAGAAAAAACATAAATGTAATCCAAAATGCAATCATATTCATATAAATTAACA  
TGTTGTTATGTACATGATAATGTTATGGAGATGCAGAAAAAACATAAATGTAATCCAAAATGCAATCATATTCATATAAATTAACA  
C C Y V H D N C Y G D A E K K H K C N P K T Q S Y S Y K L T
```

```
K R T I S A M Y P Q V L V H V L S V I V T A R Q P S A S A I  
AAAAGAAACACATCGCTATGGTGGCGCAGGTACTTGTGACAGTATTGCTGTGATGTGACCGCAGCGCAGCCCTGCTTCGGCGAT  
AAAAGAAACACATCGCTATGGTGGCGCAGGTACTTGTGACAGTATTGCTGTGATGTGACCGCAGCGCAGCCCTGCTTCGGCGAT  
K R T I I C Y G A A G T C A R I V C D C D R T A A L C F G D
```

```
L N T S S G T R I L T F R D I A  
TCFGAATAGATGACGGCGCACAAGAATATTGACCGCGGAGATATGGC  
TCFGAATACATCGAGCGGCACAAGAATATTGACACCGCGAGATTGGC  
S E Y I E R H K N I D T A R F C
```



# Mammals Platelet-derived growth factors

*Homo Sapiens & Rattus norvegicus*

# Mammals Platelet-derived growth factors

*Homo Sapiens & Rattus norvegicus*

PDGFA_HUMAN	1	MRTLACLLLLGCGYLAHVLAEEAEIPREVIERLARSQIHSIRDLQRLL	50
BAA00987.1	1	MRTWACLLLLGCGYLAHALAEEAEIPRELIERLARSQIHSIRDLQRLL	50
PDGFA_HUMAN	51	DSVGSSELDTSLRAHGVHATKXVPEKRPLPIRRKRSIEEAVPAVCKTRT	100
BAA00987.1	51	DSVGAEDALETNLRAGSHTVKXVPEKRPVPIRRREVLKPFQFARPGR	100
PDGFA_HUMAN	101	VIYEIPRSQVDPTSANFLIWPPCVEVKRCTGCCNTSSVKCQPSRVHHRV	150
BAA00987.1	101	SFTRYLGARWTPTSANFLIWPPCVEVKRCTGCCNTSSVKCQPSRVHHRV	150
PDGFA_HUMAN	151	KVAKVEYVRKKPKLKEVQVRLEEHLEACATTSNLPDYREEDT	193
BAA00987.1	151	KVAKVEYVRKKPKLKEVQVRLEEHLEACATSNLNPDRHEET	193

# Mammals Platelet-derived growth factors

*Homo Sapiens* & *Rattus norvegicus*

M R T L A C C L L L L G C G V Y L A H V L A E E A E I P R E V I  
ATGAGAACATTGGCATGTTTATTATTATTAGGATGGGATATTTAGCACATGGTTAGCAGAAAGAGCAGAAAATACCAAGAGAAAGTGATA  
ATGAGAACATGGGCATGTTTATTATTATTAGGATGGGATATTTAGCACATCGCTTAGCAGAAAGAGCAGAAAATACCAAGAGAACATGATA  
M R T W A C L L L L L G C G V Y L A H A L A E E A E I P R E L I

E R L A R S Q I H S I R D L Q R L L E I D S V G C S E D S L D  
GAAAGATTAGCAGATCACAAAATACATCAATAAGAGATTTACAAGATTTATAGAAAATAGATTCAGTAGGATCGGAGATTCGTTAGAT  
GAAAGATTAGCAGATCACAAAATACATCAATAAGAGATTTACAAGATTTATAGAAAATAGATTCAGTAGGAGCCGGAAGATTCGGTTAGAA  
K R L A R S Q I H S I R D L Q R L L E I D S V G A E D A L E

T S L R A H G V Y H A T K H V P E K K R P L P I R R K R S I E E  
ACAAGTTTAAAGAGCACATGGAGTGCATGCGACGAAACATGTACCAGAAAAAAGACCCTGCCAATAAGAAGAAAGAGAAATTTGAGGAA  
ACAAATTTAAGACACATGGATCGCATACGGTGAACATGTACCAGAAAAAAGACCAGTCCCAATAAGAAGAGAGAGAAAGTATTGAGGAA  
T N L R A H G S H T V K H V P E K R P V P I R R R E R V L R K

A V P A V C K T R T V I V E I P R S Q V D P T S A N F L I W  
GCCGTTCCCGCAGTTTGGCAAGACGAGGAGGGTCATTTACAGATACCTAGGAGCCAGGTGGACSCAACATCAGGAAAATTTTTTAAATATG  
CCGTTCCCGCAGTTTGGCAAGACGAGGAGGGTCATTTACAGATACCTAGGAGCCAGGTGGACSCAACATCAGGAAAATTTTTTAAATATG  
P F P Q F A R P G R S F T R Y L G A R W T P T S A N F L I W

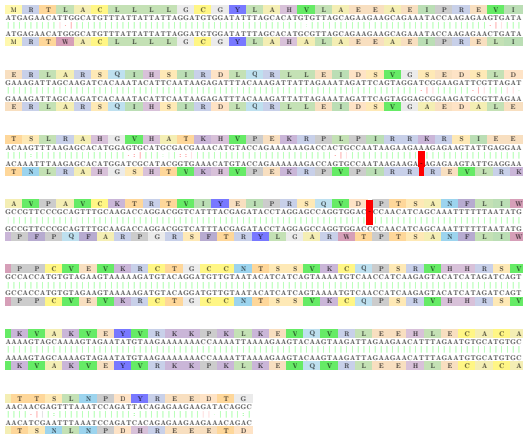
F P C V E A V K R C T G C C N T S S V K C Q P S R V B H R S V  
GCCACCATGTGTAGAAGTAAAGAGATGTACAGGATGTGTAAATACATCATCAGTAAAATGTCAACCATCAAGAGTACATCATAGATCAAT  
GCCACCATGTGTAGAAGTAAAGAGATGTACAGGATGTGTAAATACATCATCAGTAAAATGTCAACCATCAAGAGTACATCATAGATCAAT  
P F P C V E V K R C T G C C N T S S V K C Q P S R V B H R S V

K V A K V E Y V R K K P K L K E V Q V R L E E H L E C A C A  
AAAAGTAGCAAAAGTAGAATAGTAAAGAAAAAACCAAAATAAAAGAGTACAAGTAAAGATTAGAAGAACATTTAGAATGTGCATGTC  
AAAAGTAGCAAAAGTAGAATAGTAAAGAAAAAACCAAAATAAAAGAGTACAAGTAAAGATTAGAAGAACATTTAGAATGTGCATGTC  
K V A K V E Y V R K K P K L K E V Q V R L E E H L E C A C A

T T S L N P D Y R E E D T G  
AACACGAGTTTAAATCCAGATTACAGAGAAGAAGATACAGGC  
AACATCGAATTTAAATCCAGATCACAGAGAAGAAGAACAGAC  
T S N L N P D H R R E E E T D

# 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*

```
M R T L A C L L L L L G C G Y L A H V L A E E A E I P R E V I
ATGAGAACATGGCCATGTTTATTATTATTAGGATGGGATATTTAGCACATGGTTAGCAGAAAGAGCAGAAAATACCAAGAGAAAGTGTATA
ATGAGAACATGGCCATGTTTATTATTATTAGGATGGGATATTTAGCACATGGTTAGCAGAAAGAGCAGAAAATACCAAGAGAAATCGTAT
M R T W A C L L L L L G C G Y L A H A L A E E A E I P R E L I
E R L A R S Q I H S I R D L Q R L L E I D S V G C S E D S L D
GAAAGATFAGCAGATCACAAAATACATCAATAAGAGATTTACAAGATTTATAGAAAATAGATTCAGTAGGATCGGAGATTCGTTAGAT
GAAAGATFAGCAGATCACAAAATACATCAATAAGAGATTTACAAGATTTATAGAAAATAGATTCAGTAGGAGCCGGAAGATCGGTTAGAA
K R L A R S Q I H S I R D L Q R L L E I D S V G A E D A L E
T S L R A H G Y H A T K H V P E K K R P L P I R R K R S I E E
ACAAGTTTAAAGAGCATTGGAGTGCATGCCAGAAAACATGTACCAGAAAAAGACCCTGCCAATAAGAAAGAAAGAGAAAGTATTGAGGAA
ACAAATTTAAGAGCATTGGATCGCATACGGTGAACATGTACCAGAAAAAGACCCTGCCAATAAGAAAGAAAGAGAAAGTATTGAGGAA
C N L R A H G S H T Y K H V P E K R P V P I R R R E R E V L R K
A V F A V C K T R T V I V E I P R S Q V D P T S A N F L I W
GCCGTTCCCGCAGTTTGGCAAGACGAGGAGGGTCATTTACGAGATACCTAGGAGCCAGGTGGACSCAACATCAGCAAAATTTTTTAAATATG
GCCACCATGTGTAGAAAGTAAAAGATGTACAGGATCTGTAAATACATCATCAGTAAAATGTCAACCATCAAGAGTACATCATAGATCAAT
CCGTTCCCGCAGTTTGGAAAGACGAGCGGTCATTTACGAGATACCTAGGAGCCAGGTGGACCCAACATCAGCAAAATTTTTTAAATATG
P F P Q F A R P G R S F T R Y L G A R W T P T S A N F L I W
F P C V E A V K R C T G C C N T S S V K C Q P S R V B H R S V
GCCACCATGTGTAGAAAGTAAAAGATGTACAGGATCTGTAAATACATCATCAGTAAAATGTCAACCATCAAGAGTACATCATAGATCAAT
P F C V E A V K R C T G C C N T S S V K C Q P S R V B H R S V
K V A K V E Y V R K K P K L K E V Q V R L E E H L E C A C A
AAAAGTAGCAAAAGTAGAATAFTGAAGAAAAAAACAAAATTAAGAAGTACAAGTAAAGATTAGAAGAACATTTAGAATGTCATGATCGC
AAAAGTAGCAAAAGTAGAATAFTGAAGAAAAAAACAAAATTAAGAAGTACAAGTAAAGATTAGAAGAACATTTAGAATGTCATGATCGC
K V A K V E Y V R K K P K L K E V Q V R L E E H L E C A C A
T T S L N P D Y R E E D T G
AACACGAGTTTAAATCCAGATTACAGAGAAGAAGATACAGGC
AACATCGAATTTAAATCCAGATCACAGAGAAGAAGAACAAGAC
T S N L N P D H R R E E E T D
```


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)



# Proposed tool

Available at <http://bioinfo.lifl.fr/path/>



## path :: web server

home   web server   examples   retrieve result with an ID

**bioinfo.lifl.fr**  
Sequoia

**mreps**

**YASS**

**Path**

**Magnolia**

**Protea**

**Reblosium**

**RNA**  
Carnac  
RNAfamily  
Gardenia  
REGLISS

**TFM**  
TFM-Explorer  
TFM-Scan  
TFM-Pvalue

**Norine**  
General  
Structure  
Monomers

### SEQUENCES

**Protein sequence(s) 1**

Paste one sequence in FASTA format:

```
>sp|Q8AY47|PA2A2_BUNCA  
AVCVLLGANKPPSPFHLNIPKRNINYYIPCKEYHGEYVQCYCGVGGSPFDALDR  
CCYVNDYDGAEEKKCNPNQSTYKLEKHTISAKYQVLVHVLVSVIARQPSAGAI  
LNTSGSTRLTFRDIADDI
```

or

Upload MULTIFASTA file:

**Protein sequence(s) 2**

Paste one sequence in FASTA format:

```
>sp|P06503|PA22_NA2MO  
NLYGFENIRCTYDSEPKHAYDSCYQGRGRGTAVDGLDRCCQVSDNICYGAEKLCW  
PYLLTKYECGGKLTCSGGHREKAAAVCNCLVAANCYAGARYIDANYINLKERCO
```

or

Upload MULTIFASTA file:

[Sequence examples >](#)

### PARAMETERS

**Scoring function**

**Translation-dependent base substitution scores**  
Assumed evolutionary distance between sequences:  mutations per codon (more information about the translation dependent scores)

**Classic base substitution scores**  
Match: 03   Transversion mutation: -04   Transition mutation: -02

**Gap penalties**

**Gap of size 3** (codon insertion or deletion):  
opening: -30   extension: -10

**Frameshift gap** (insertion/deletion of one or two bases):  
opening: -60   extension: -20

**Alignment sense**

**Forward** (align the first sequence with the second sequence, both in the forward translation sense)

**Backward** (align the first sequence with the reverse complementary of the second sequence)

**Forward and Backward**



# 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

# 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 ...)

Thank you for your attention

<http://bioinfo.lifl.fr/path/>

# References I



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, Stocholm, 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.

# References II



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.

# References III



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.