Federal University of Rio de Janeiro
Institute of Biophysics Carlos Chagas Filho
Bioinformatics for Transcriptomics and Functional Genomics

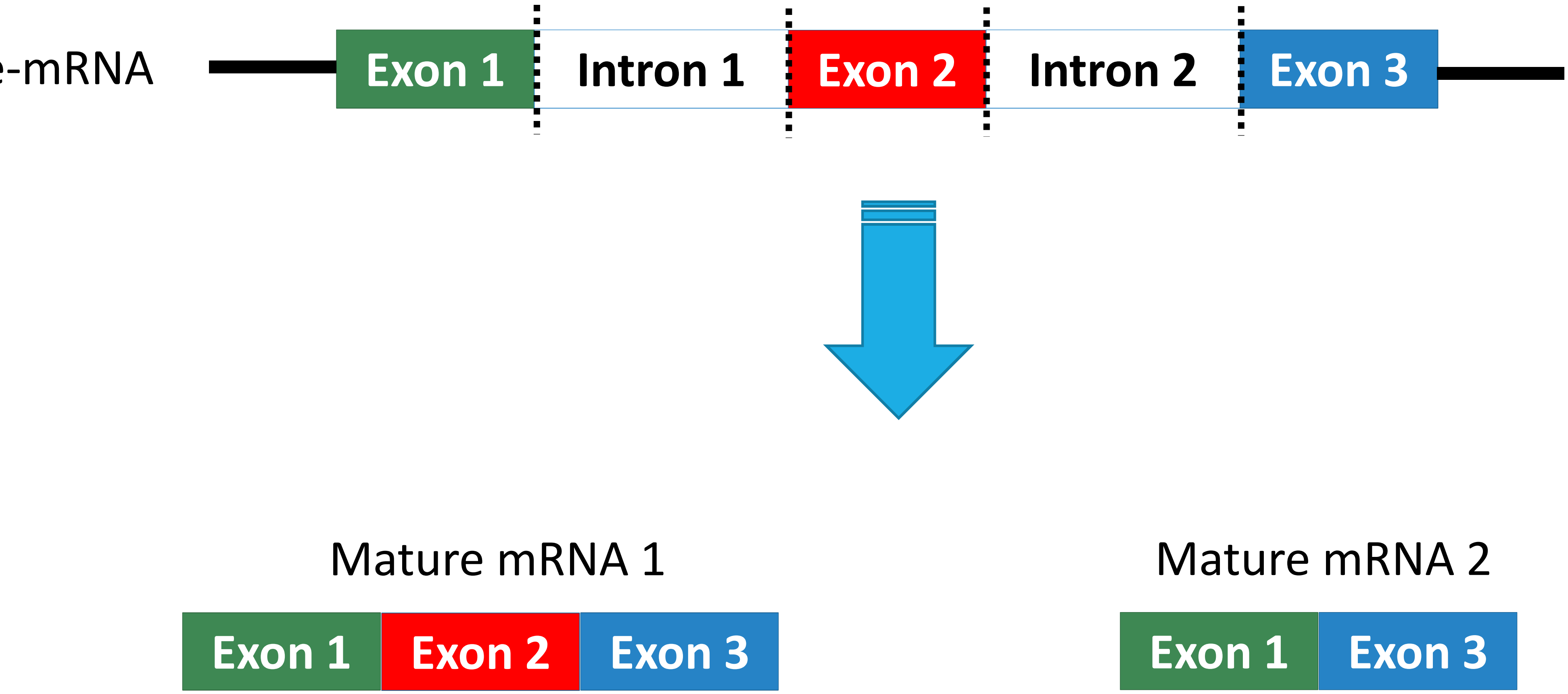
tomated Functional Annotation of Pro

Products of Alternatively Spliced Gene

Vitor Coelho & Michael Sammeth

February 2018, Bled Winterseminar

Alternative Splicing (AS)



e-mRNA

Exon 1 Intron 1 Exon 2 Intron 2 Exon 3

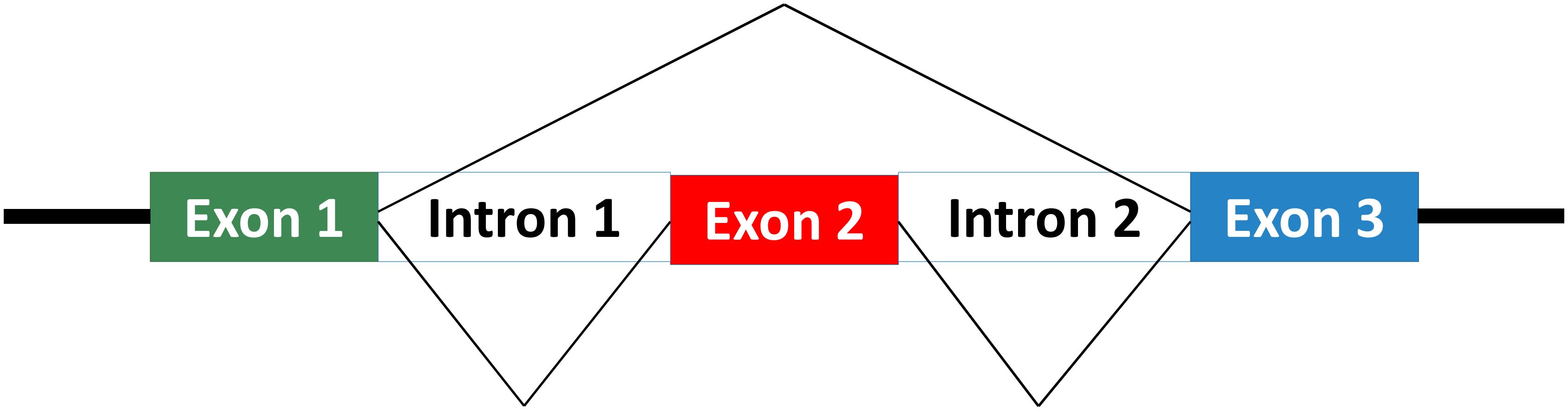
Mature mRNA 1

Exon 1 Exon 2 Exon 3

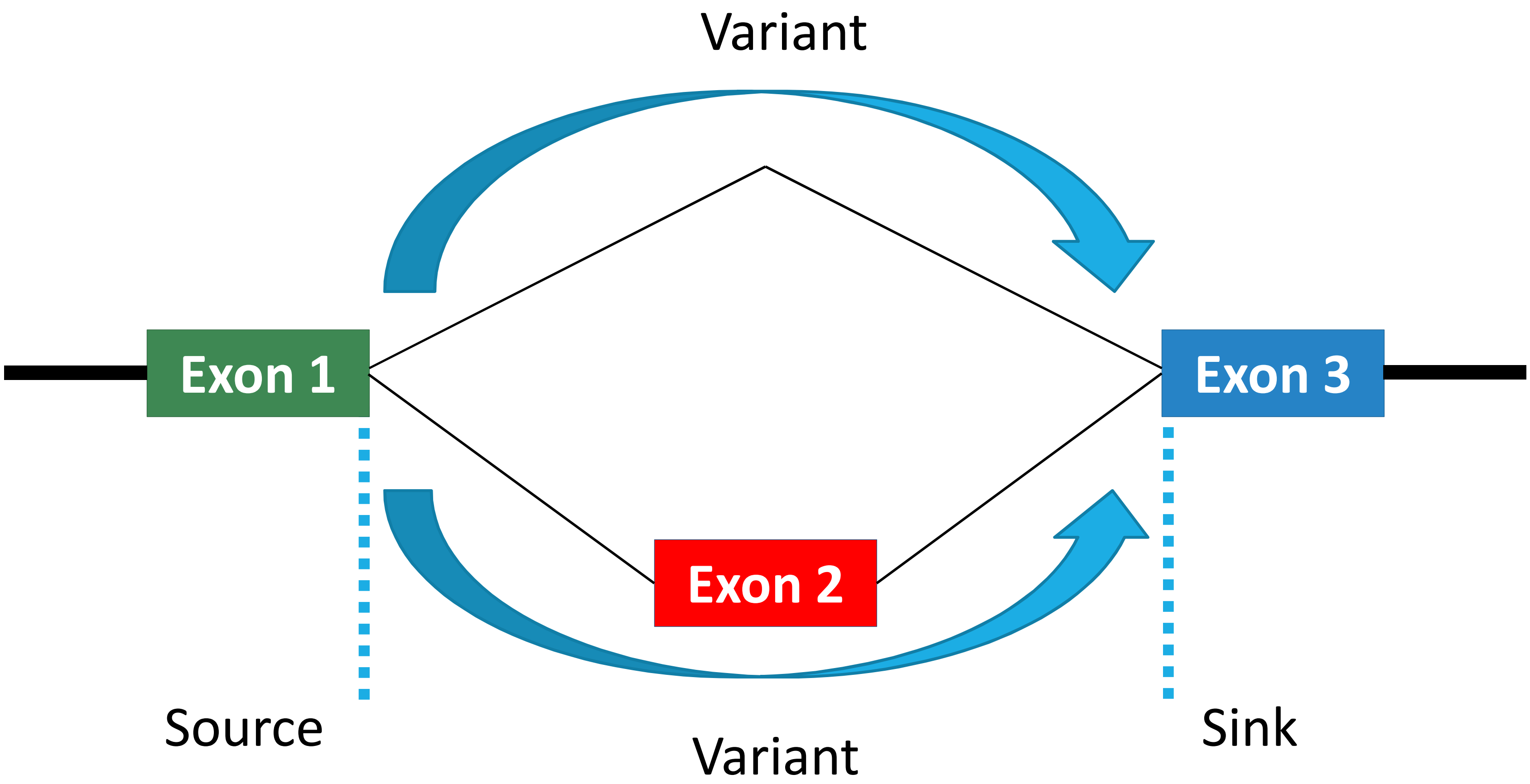
Mature mRNA 2

Exon 1 Exon 3

Alternative Splicing Events



Components of an AS Event



[AStalavista - Foissac&Sammeth 20

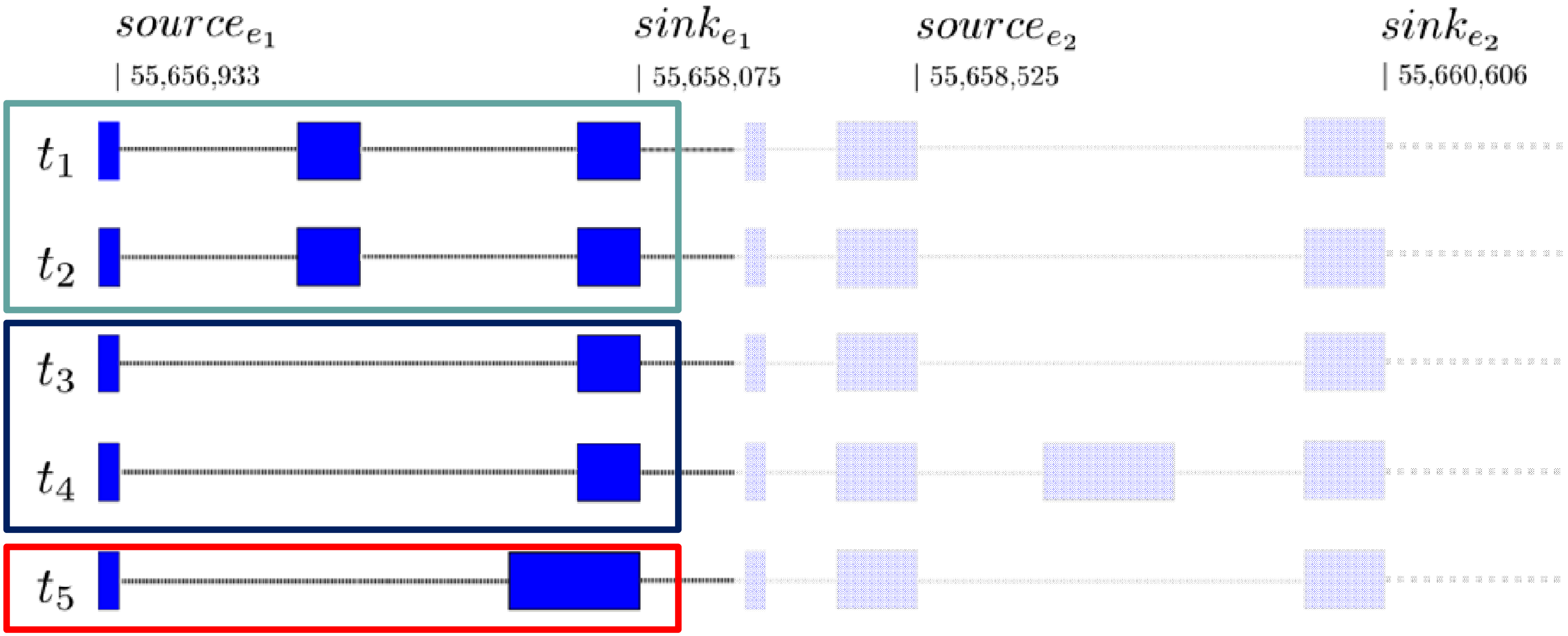
Definition of AS event

- An event e is defined by

$$e = (\, source_e, sink_e, \mathbb{V}_e)$$

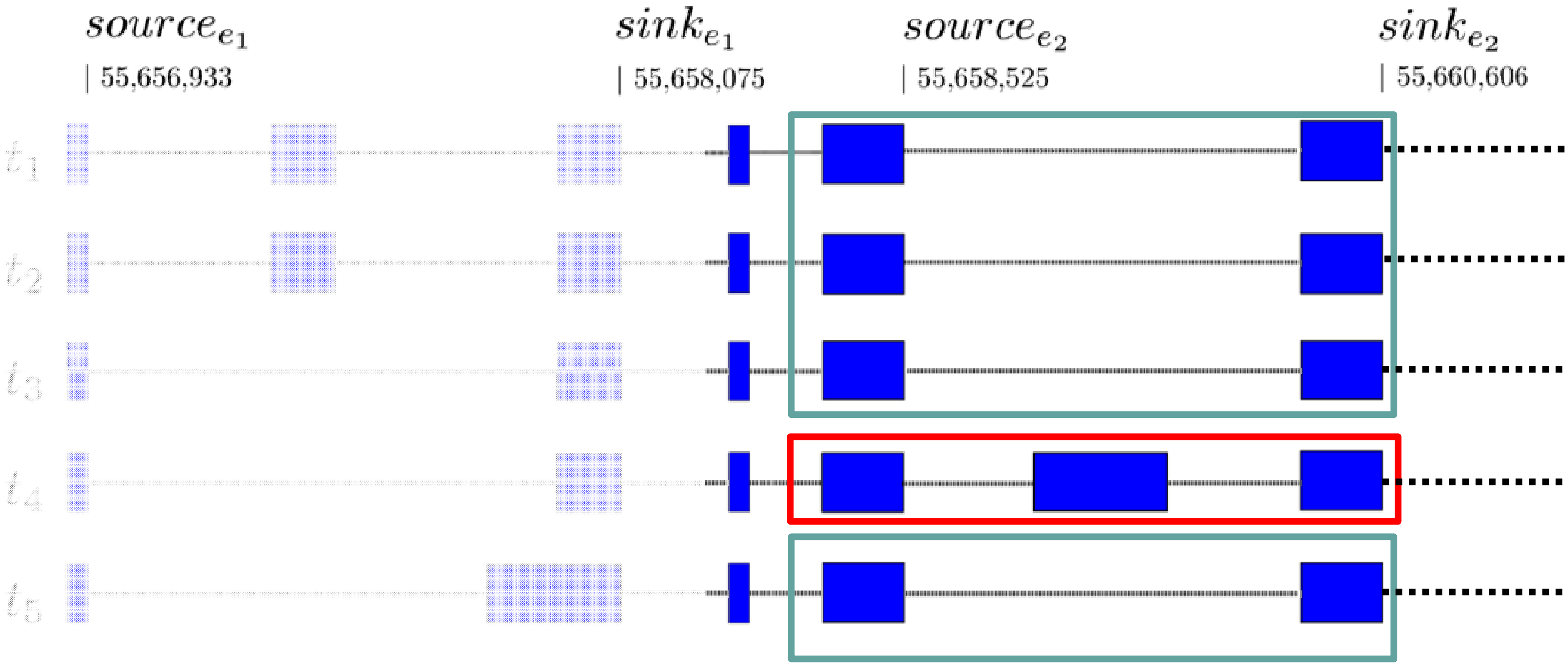
  - Where:
    - *source*$_e$: begin coordinate of event e;
    - *sink*$_e$: end coordinate of event e;
    - $\mathbb{V}_e$: set of variants;
      - A variant describes an exon/intron structure between *source*$_e$ and *sink*$_e$ – which can contain multiple transcripts

Event 1 in TNNT1



$$\mathbb{V} = \{V_1, V_2, V_3\} \rightarrow \begin{array}{l} V_1 = \{t_1, t_2\} \\ V_2 = \{t_3, t_4\} \\ V_3 = \{t_5\} \end{array}$$

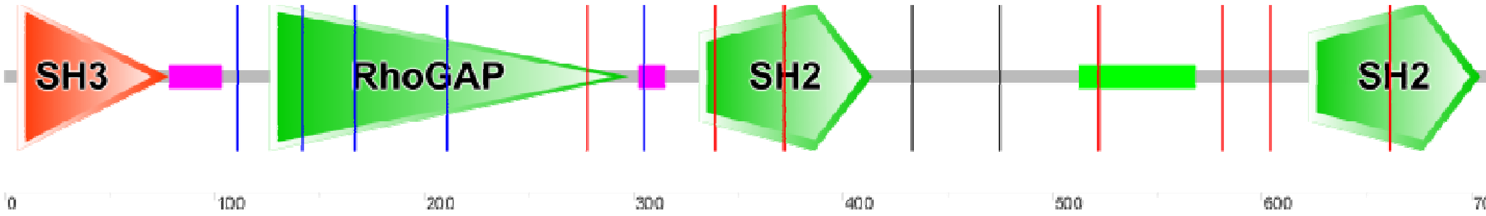
Event 2 in TNNT1



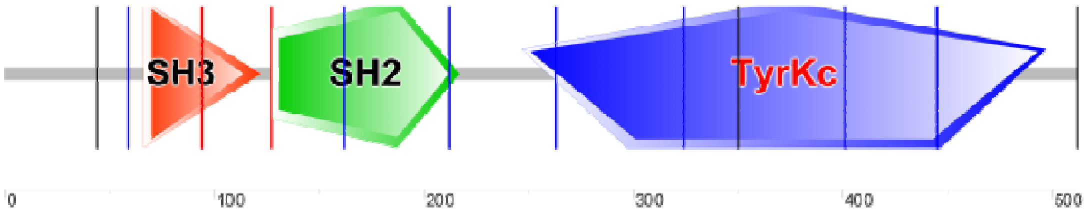
$$\mathbb{V} = \{\mathrm{V}_1, \mathrm{V}_2\} \rightarrow \begin{array}{l} V_1 = \{t_1,t_2,t_3 \\ V_2 = \{t_4\} \end{array}$$

Protein Domains and Splicing

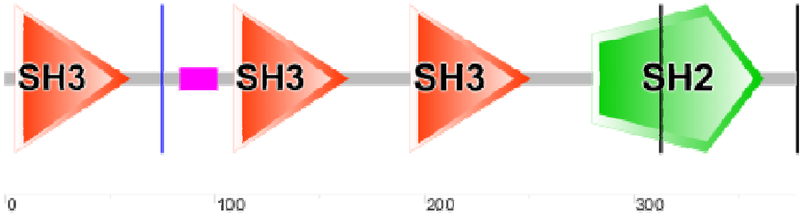
Human Protein p85A



Human Protein Lyn



Human Protein Nck1



The Pfam Database

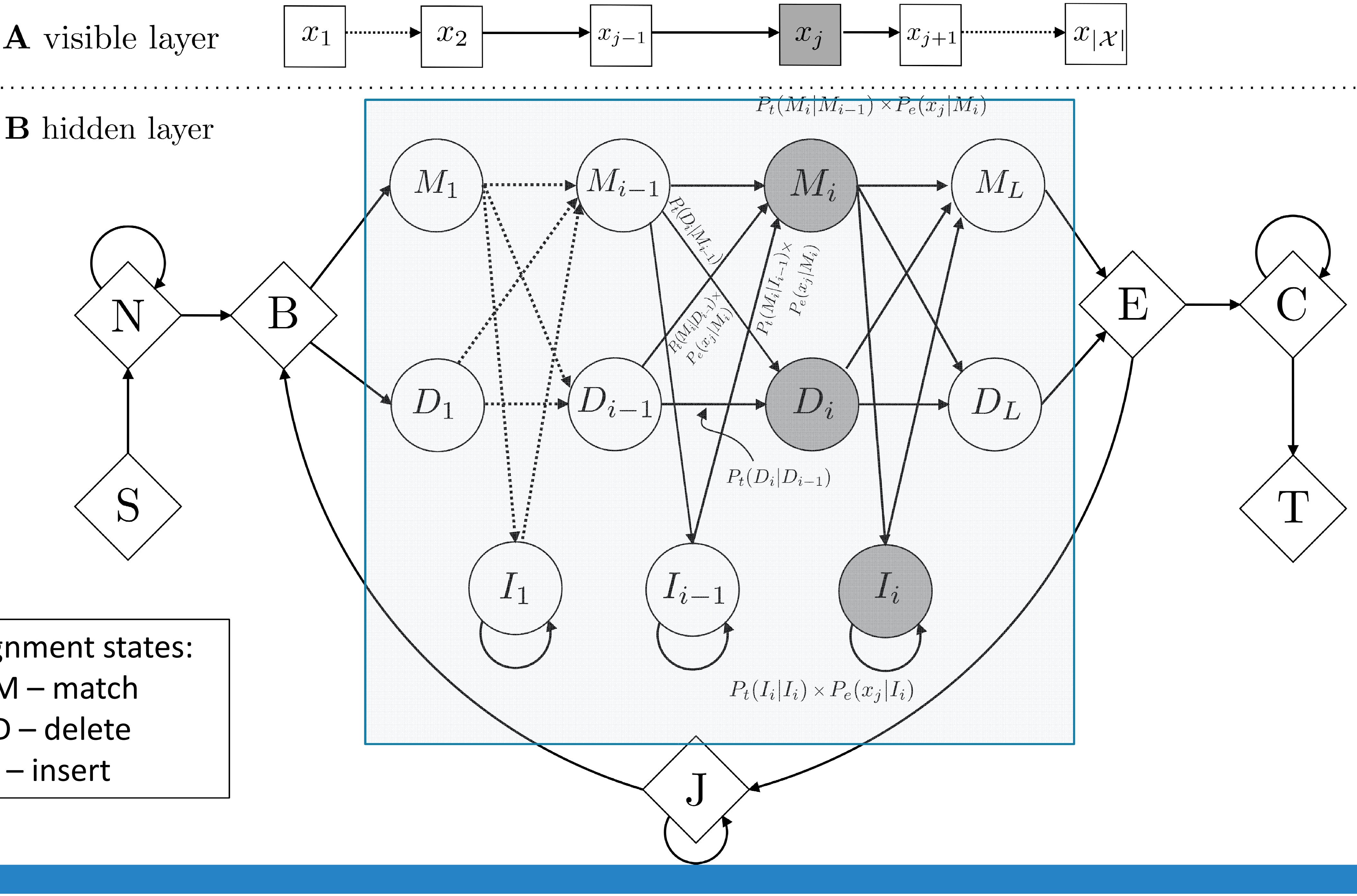
- *Protein fam*ilies = database of protein families [Punta et al, 2012]

- Pfam domains are organized in different sub-groups:

  Pfam-A: high quality and manually curated protein families

  Pfam-B: automatic predictions

- protein regions sharing sequence similarity of high significance θ_{π^i} (*“domains”*) are represented by profile HMMs Π , each with a specific length L_Π, a prediction threshold θ, and alignment scores (amino acid scores and gap penalties)



gnment states:
M – match
D – delete
– insert

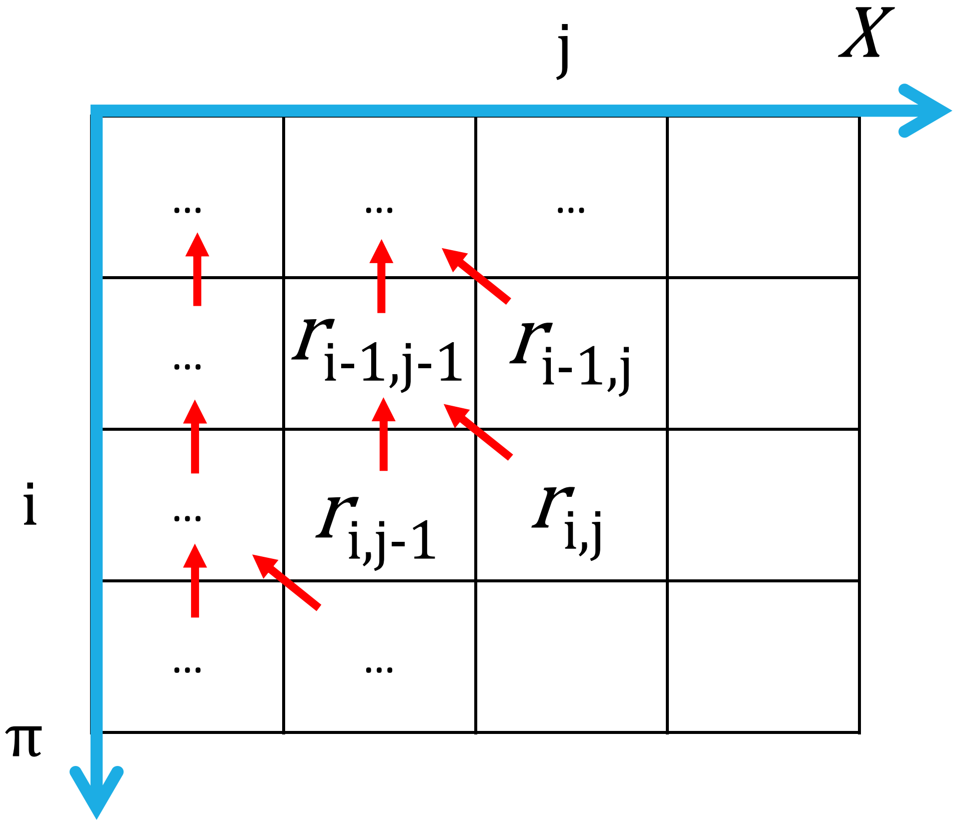
Pfam Profile HMM file format

```
HMMER3/b [3.0 | March 2010]
NAME  Antimicrobial11
ACC   PF08106.6
DESC  Formaecin family
LENG  16
...
GA    25.00 25.00;
...
HMM         A         C         D     ...    S          T         V         W         Y
          m->m      m->i      m->d          i->m       i->i      d->m      d->d
 COMPO   3.68653   5.37452   4.20831  ...  3.83085    2.62022   2.91489   5.77804   3.54633
         2.68618   4.42225   2.77519  ...  2.37887    2.77519   2.98518   4.58477   3.61503
         0.01467   4.62483   5.34718      0.61958     0.77255   0.00000      *
   1     3.84125   5.47999   4.65486  ...  4.04240    4.37307   4.91201   6.51694   5.82696    1 - -
         2.68618   4.42225   2.77519  ... 2.37887     2.77519   2.98518   4.58477   3.61503
          0.01467  4.62483   5.34718      0.61958     0.77255   0.48576   0.95510
   2     4.09153   5.75008   4.67229  ...  4.17197    4.33362   4.67489   6.11461   5.14175    2 - -
         2.68618   4.42225   2.77519  ... 2.37887     2.77519   2.98518   4.58477   3.61503
         0.01467   4.62483   5.34718      0.61958     0.77255   0.48576   0.95510
     ...
  15     4.09153   5.75008   4.67229  ...  4.17197    4.33362   4.67489   6.11461   5.14175    15 - -
         2.68618   4.42225   2.77519  ...  2.37887    2.77519   2.98518   4.58477   3.61503
         0.01467   4.62483   5.34718      0.61958     0.77255   0.48576   0.95510
  16     4.12723   5.39816   5.40081  ...  4.69892    4.40337   3.12238   5.70016   4.58094    16 - -
         2.68618   4.42225   2.77519  ...  2.37887    2.77519   2.98518   4.58477   3.61503
         0.00990   4.62006      *         0.61958     0.77255   0.00000      *
```

Viterbi algorithm

- Dynamic Programming
  - recursive algorithm
  - divides exponential problem into polynomial sub-problems

- Calculate best alignment score between prefixes:
  - Profile HMM $\pi[1; i]$ and query sequence $X[1; j]$



- path of best alignment is stored by pointers to the cell that led to highest score

Viterbi algorithm

$$r^{M}_{i.j} = \log_2\left(\frac{P_e(x_j|M_i)}{P_0(x_j)}\right) + \max\begin{cases} r^{M}_{i-1.j-1} + \log_2(P_t(M_i|M_{i-1})) \\ r^{I}_{i-1.j-1} + \log_2(P_t(M_i|I_{i-1})) \\ r^{D}_{i-1.j-1} + \log_2(P_t(M_i|D_{i-1})) \end{cases}$$

$$r^{I}_{i.j} = \log_2\left(\frac{P_e(x_j|I_i)}{P_0(x_j)}\right) + \max\begin{cases} r^{M}_{i.j-1} + \log_2(P_t(I_i|M_i)) \\ r^{I}_{i.j-1} + \log_2(P_t(I_i|I_i)) \end{cases}$$

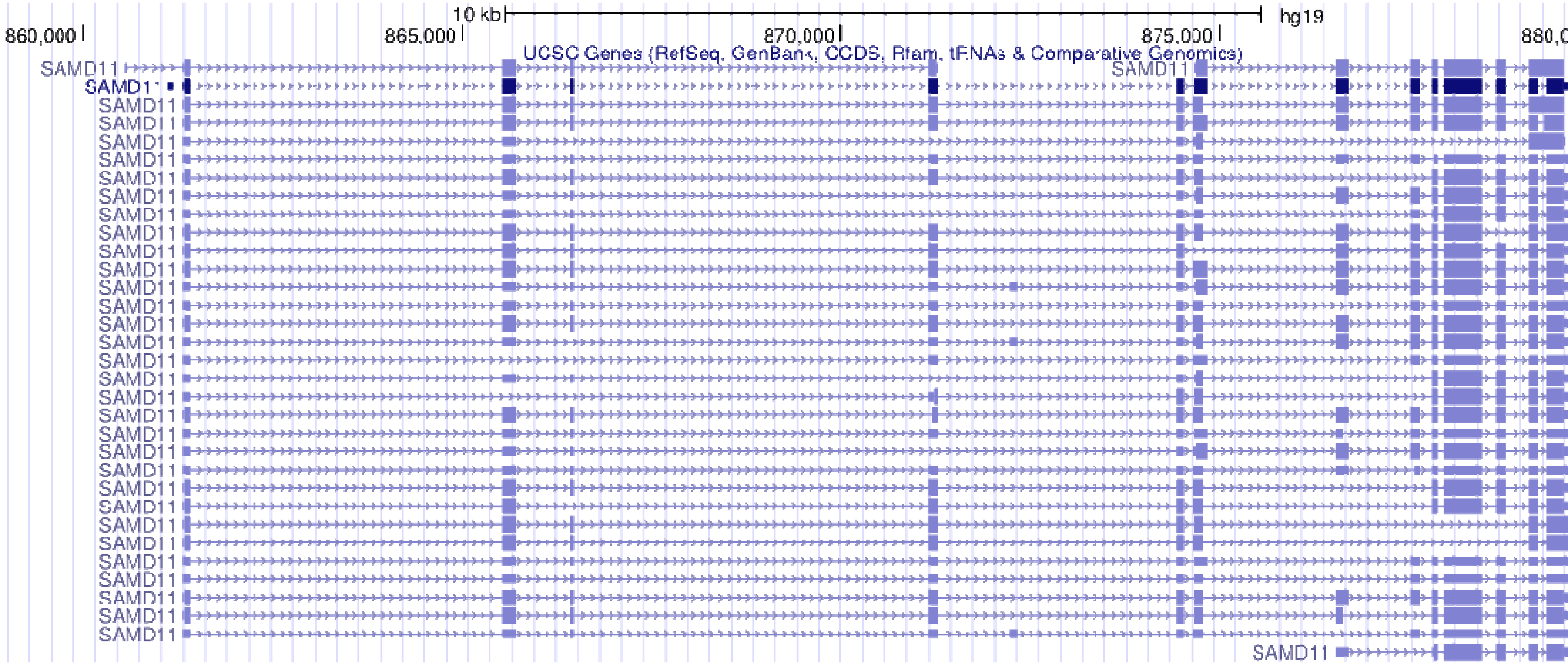
$$r^{D}_{i.j} = \max\begin{cases} r^{M}_{i.j-1} + \log_2(P_t(D_i|M_{i-1})) \\ r^{D}_{i.j-1} + \log_2(P_t(D_i|D_{i-1})) \end{cases}$$

Slide 13

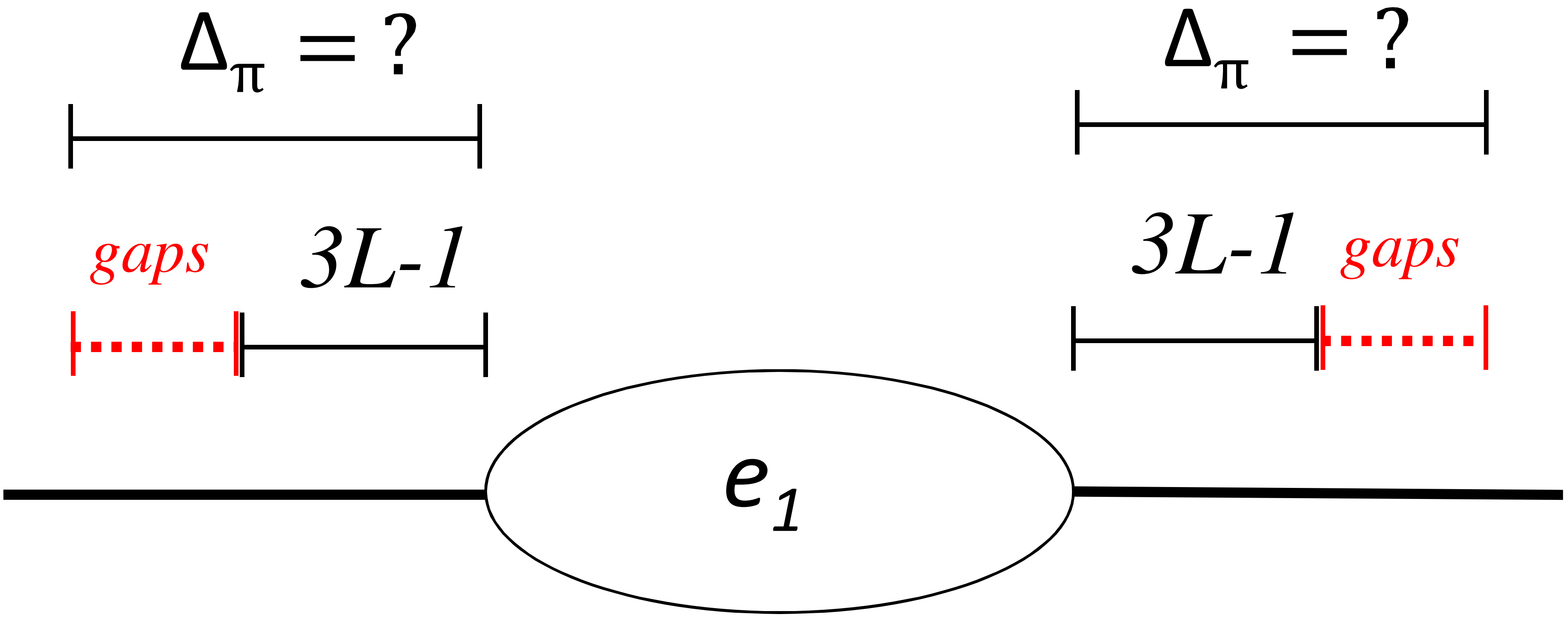
vcoelho2 está errado as probabilidades de emissao... estao como probabilidades de transicao

vitorlc; 22/2/2015

Aim: Predict Domain Alterations by AS



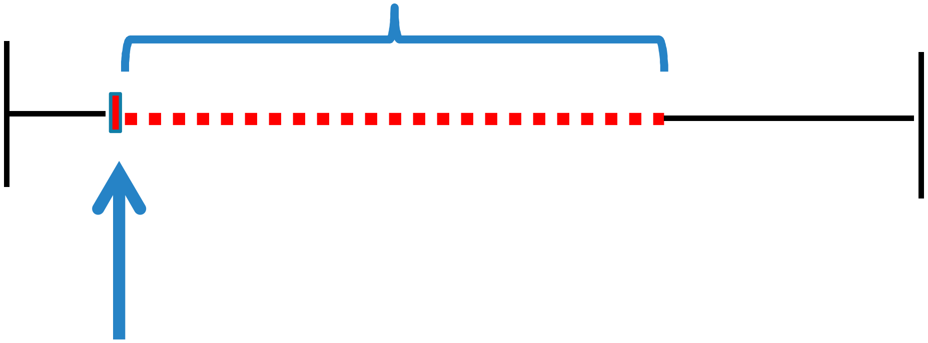
Where to scan for AS domains?



How long has to be Δ_Π?

(Gap extensions)

β β β β β β β β β



α

(Gap opening)

What is the minimum required Δ_{Π}?

$$\Delta_{\pi} = \left\lceil \frac{(L_{\pi} \times 3) - 1}{3} + \frac{\omega_{\pi}(0) - \theta_{\pi} - \alpha_{\pi}}{\beta_{\pi}} \right\rceil \times 3$$

L: *length* of profile HMM π

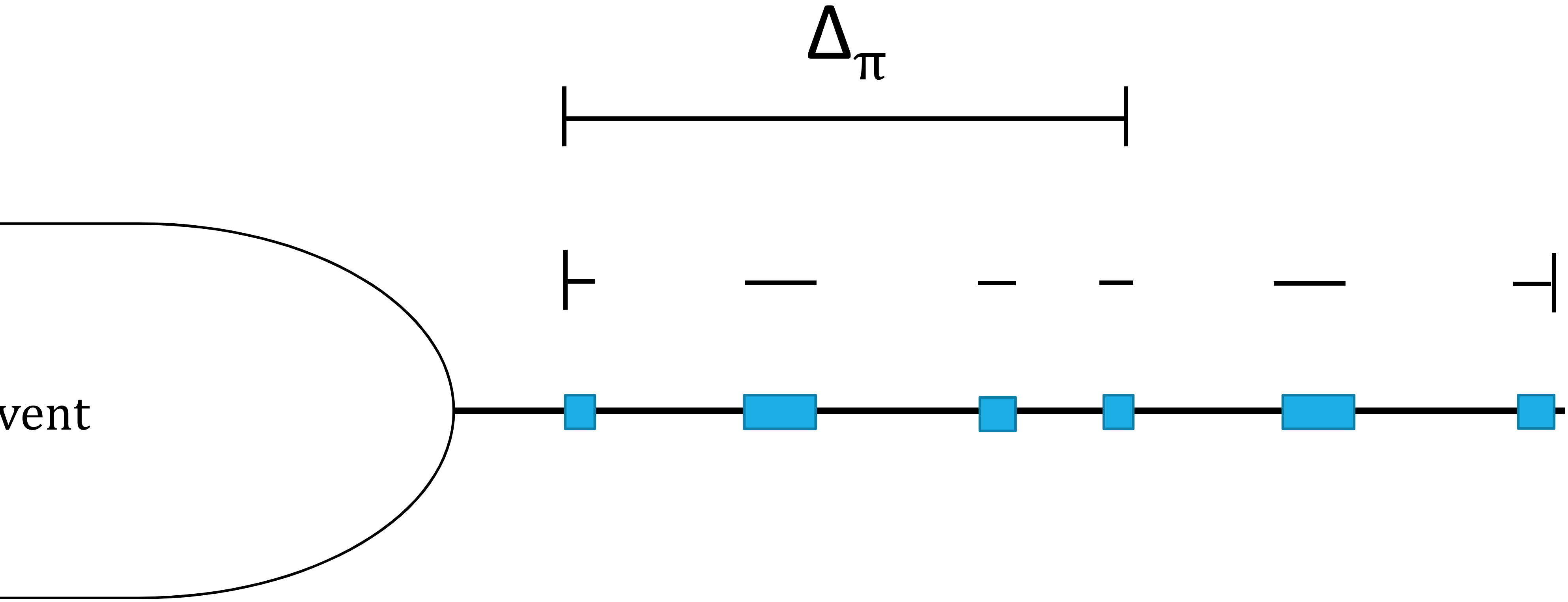
$\omega_{\pi}(1)$: *optimal alignment* score for π (optimal *suffix alignment* starting at position

θ_{π}: Domain *gathering threshold* (Pfam determined) for relevant domain score

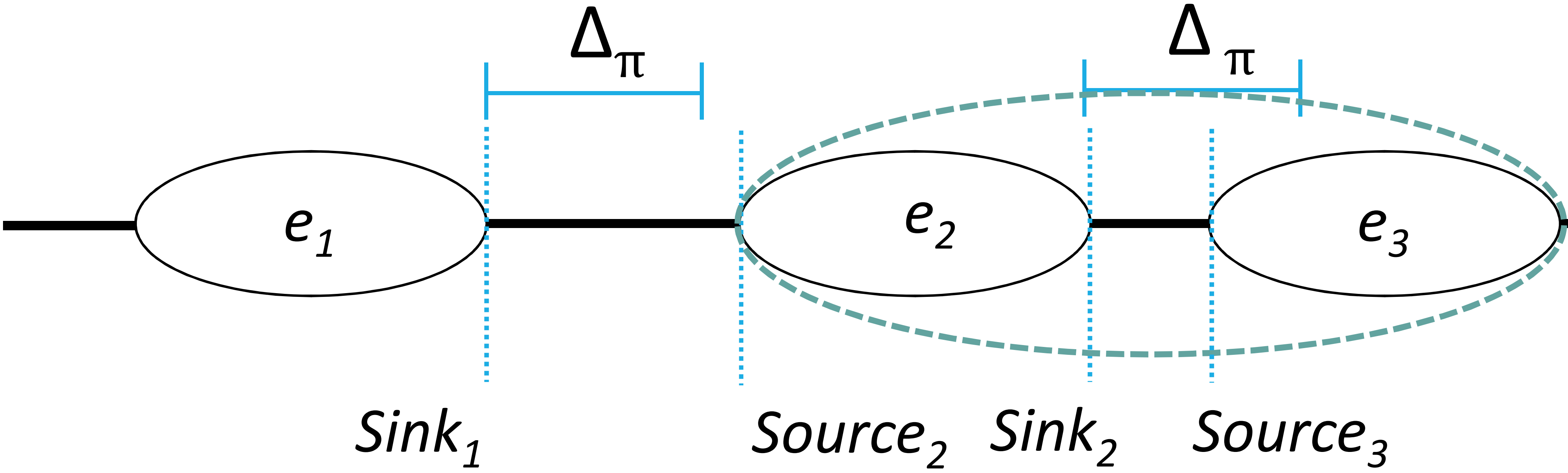
α_{π}: absolute value of max. bit score to ***open** an insertion* (“gap”, state I of the moc

β_{π}: absolute value of max. bit score to ***extend** an insertion*

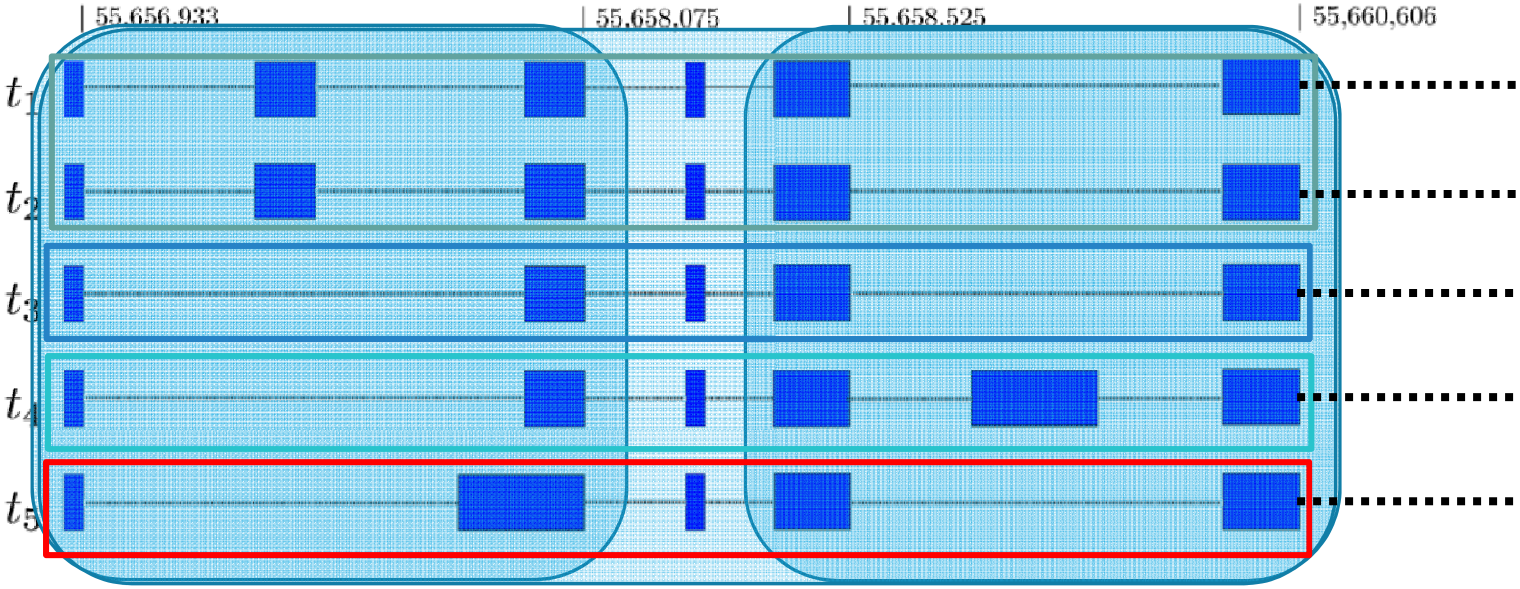
Extension of AS Event



Fusing Events



Splitting Variants



e_2

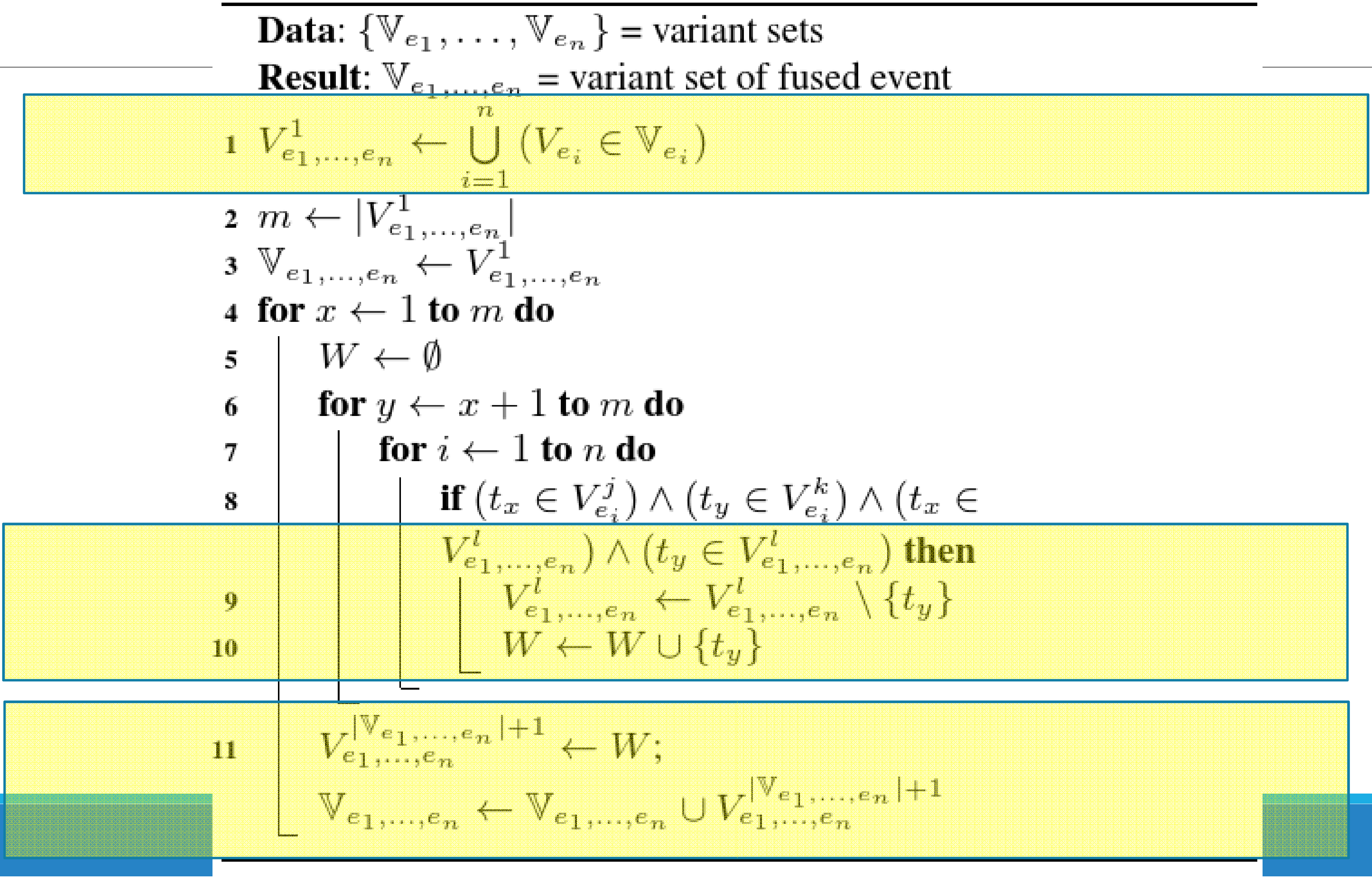
e_3

$$\left\{\begin{array}{l} V_1 = \{t_1, t_2\} \\ V_2 = \{t_3, t_4\} \\ V_3 = \{t_5\} \end{array}\right\}$$

$$\mathbb{V}_{e_3} = \left\{\begin{array}{l} V_1 = \{t_1, t_2, t_3, t_5\} \\ V_2 = \{t_4\} \end{array}\right\}$$

$$\mathbb{V}_{e_2 e_3} = \left\{\begin{array}{l} V_1 = \{t_1, t_2\} \\ V_2 = \{t_3\} \\ V_3 = \{t_4\} \\ V_4 = \{t_5\} \end{array}\right\}$$

Algorithm: Split Variants



```
Data: $\{\mathbb{V}_{e_1}, \ldots, \mathbb{V}_{e_n}\}$ = variant sets
Result: $\mathbb{V}_{e_1,\ldots,e_n}$ = variant set of fused event
$V^1_{e_1,\ldots,e_n} \leftarrow \bigcup_{i=1}^{n} (V_{e_i} \in \mathbb{V}_{e_i})$
$m \leftarrow |V^1_{e_1,\ldots,e_n}|$
$\mathbb{V}_{e_1,\ldots,e_n} \leftarrow V^1_{e_1,\ldots,e_n}$
for $x \leftarrow 1$ to $m$ do
    $W \leftarrow \emptyset$
    for $y \leftarrow x + 1$ to $m$ do
        for $i \leftarrow 1$ to $n$ do
            if $(t_x \in V^j_{e_i}) \wedge (t_y \in V^k_{e_i}) \wedge (t_x \in V^l_{e_1,\ldots,e_n}) \wedge (t_y \in V^l_{e_1,\ldots,e_n})$ then
                $V^l_{e_1,\ldots,e_n} \leftarrow V^l_{e_1,\ldots,e_n} \setminus \{t_y\}$
                $W \leftarrow W \cup \{t_y\}$
    $V^{|\mathbb{V}_{e_1,\ldots,e_n}|+1}_{e_1,\ldots,e_n} \leftarrow W;$
    $\mathbb{V}_{e_1,\ldots,e_n} \leftarrow \mathbb{V}_{e_1,\ldots,e_n} \cup V^{|\mathbb{V}_{e_1,\ldots,e_n}|+1}_{e_1,\ldots,e_n}$
```

Optimization: Branch-and-bound Condition for the Viterbi Algorithm

- improve alignment performance for finding entire domains (no partial hits).
  - reduce cells of DP matrix to calculate pruning sub-solutions that can no longer produce a relevant alignment (score ≥ θ_π).

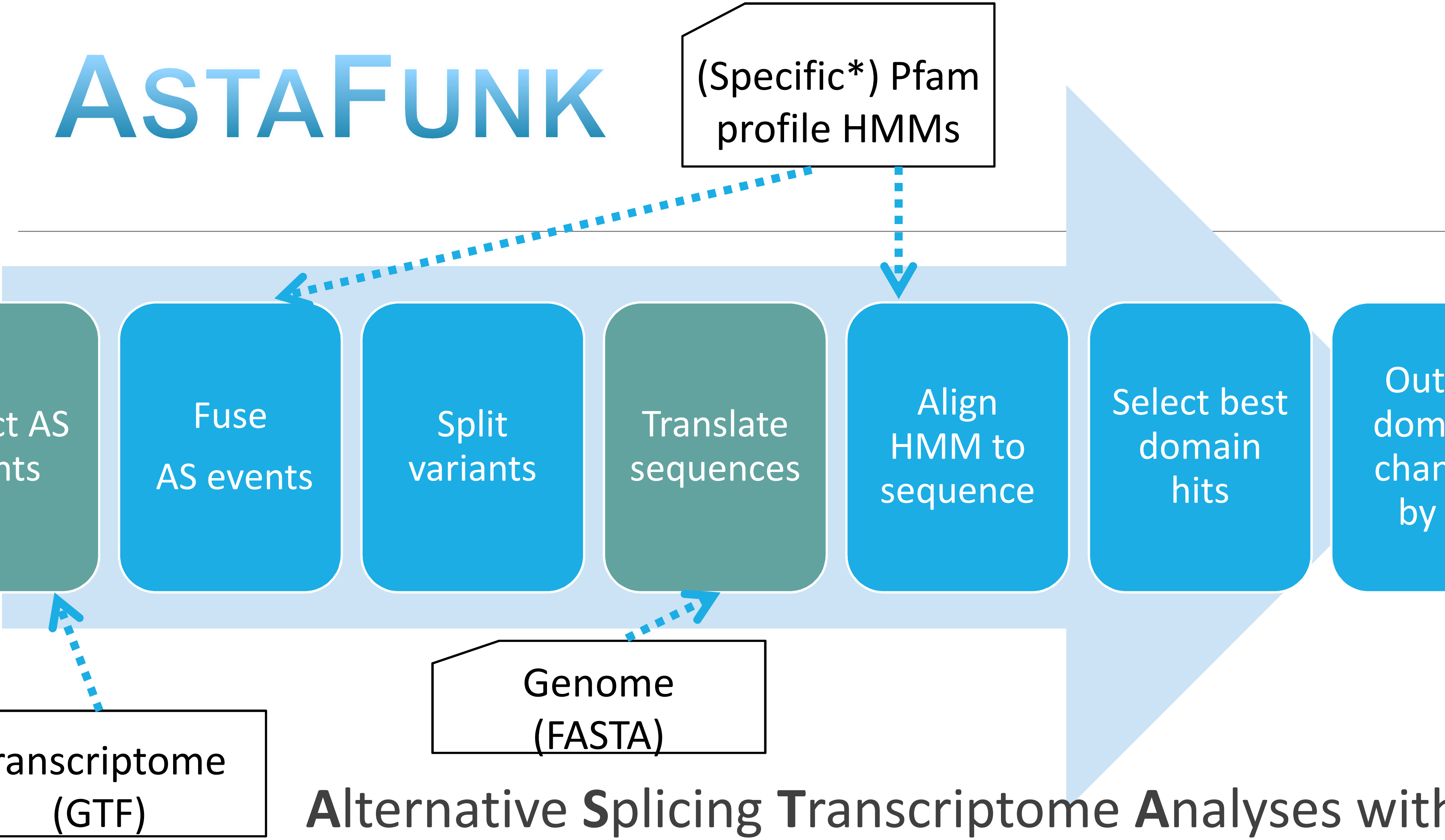
- condition for the feasibility of a sub-solution

$$r_{i,j}^{Z} + \omega(i) - \log_2(|X|) \geq \theta_\pi$$

omega = optimal suffix alignment [i..|X|]

$\log_2(|X|)$ = score normalization factor

ASTAFUNK



(Specific*) Pfam profile HMMs

ct AS nts

Fuse AS events

Split variants

Translate sequences

Align HMM to sequence

Select best domain hits

Out dom char by

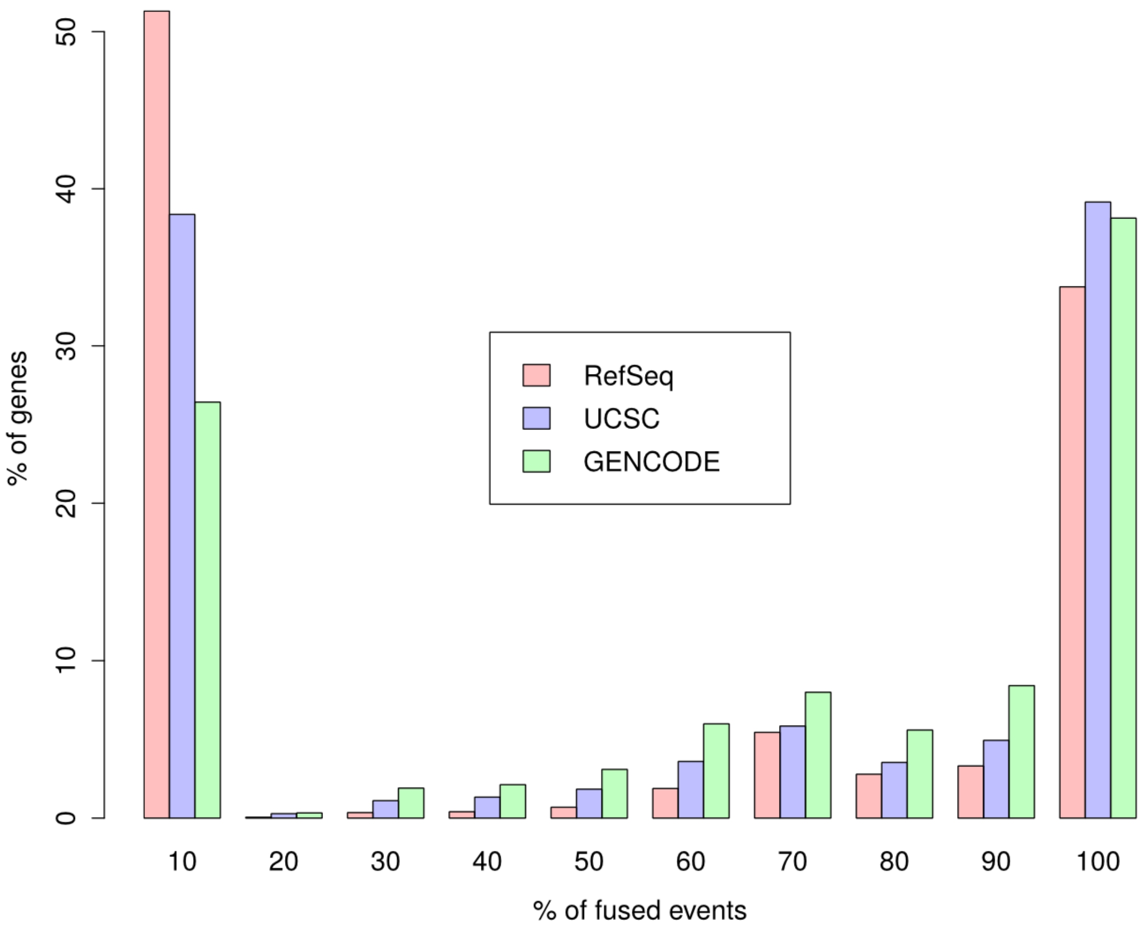
ranscriptome (GTF)

Genome (FASTA)

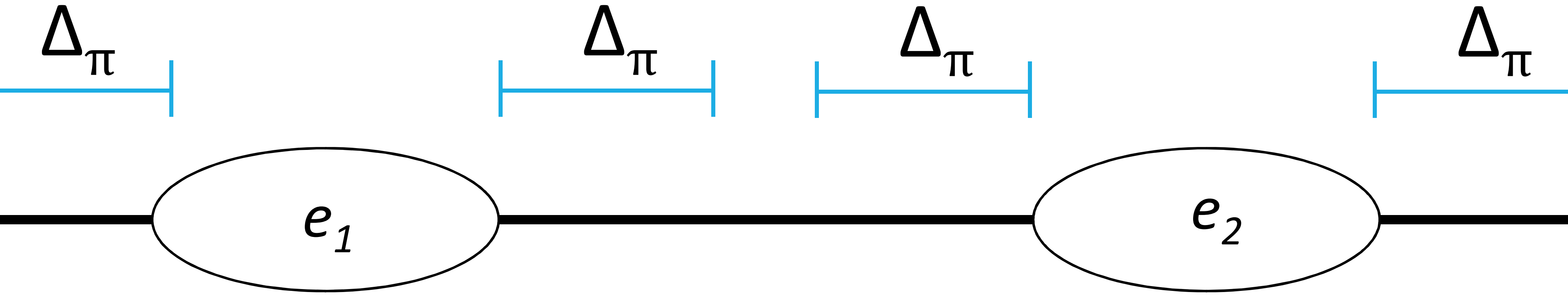
Alternative **S**plicing **T**ranscriptome **A**nalyses with **Fun**ctional **K**nowledge

	Flybase (dm3)	Wormbase (ce6)	RefSeq (hg19)	UCSC (hg19)	GEN (h
es	3,541	4,124	10,710	15,091	
es without (valid) domains	503	795	1,523	1,899	
ed AS genes	3,038	3,329	9,187	13,192	
ed AS transcripts	9,779	11,836	33,290	61,525	1
redicted domains	777	1,497	1,956	3,078	
events / AS gene	2.3	2.7	2.9	4.7	
tive search [min]	12.9	35.9	688.5	716.0	
NK search [min]	4.02	7.9	29.9	37.6	
peedup Gain	**68.84%**	**78%**	**95.7%**	**94.7%**	

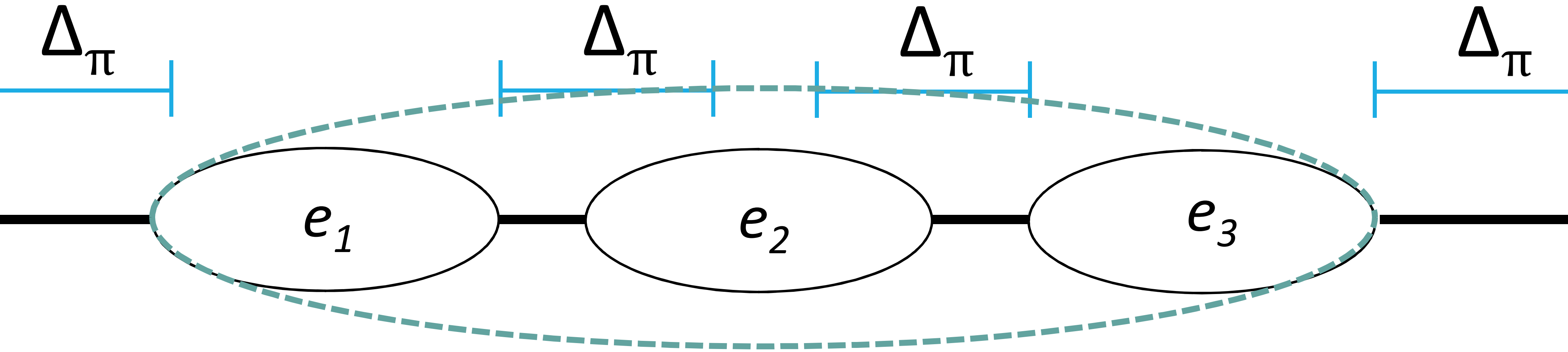
AS Density as a determinant of Fused Events (overhead)



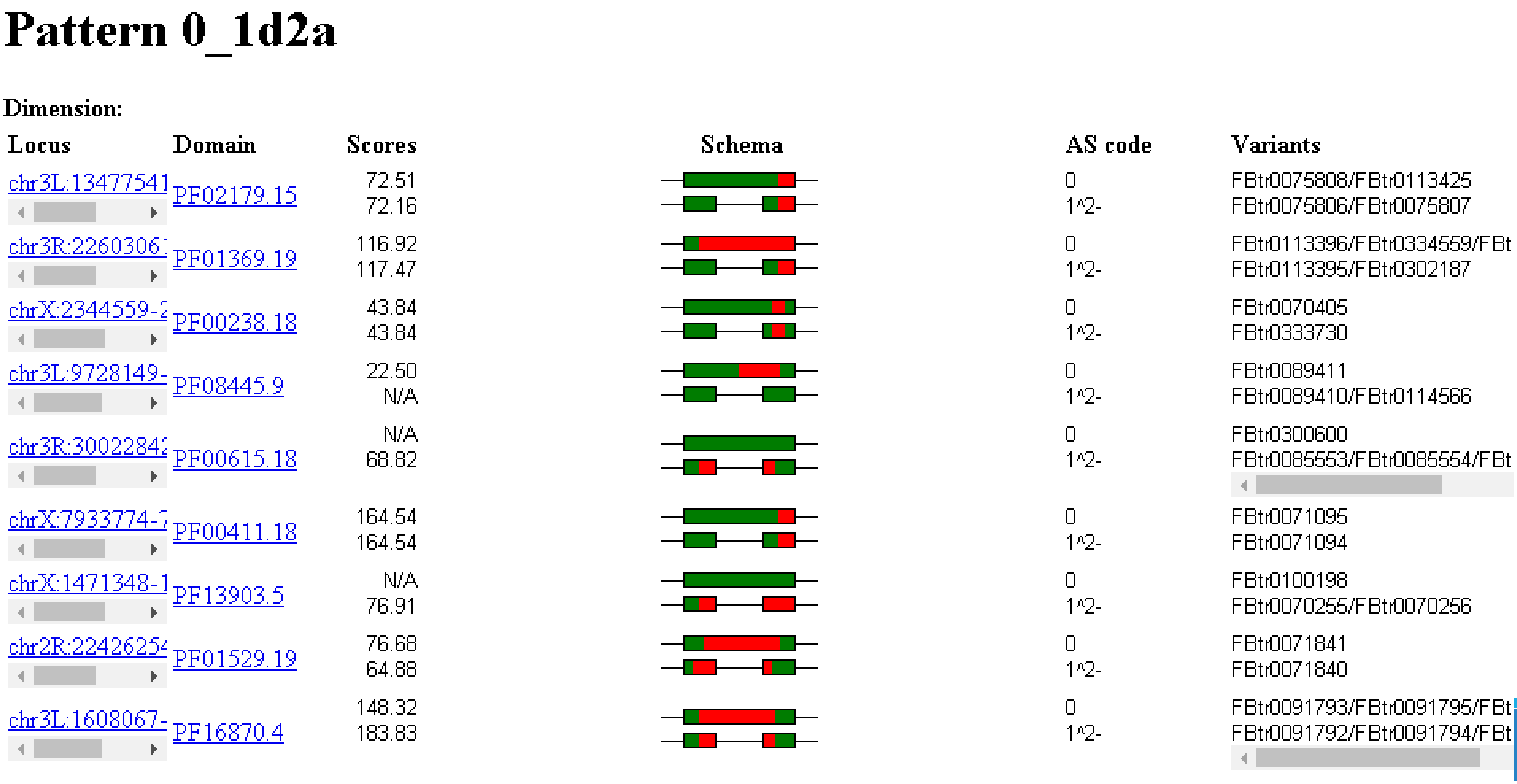
Runtime as a Function of the Density/Fusion of AS Events



Runtime as a Function of the Density of AS Events



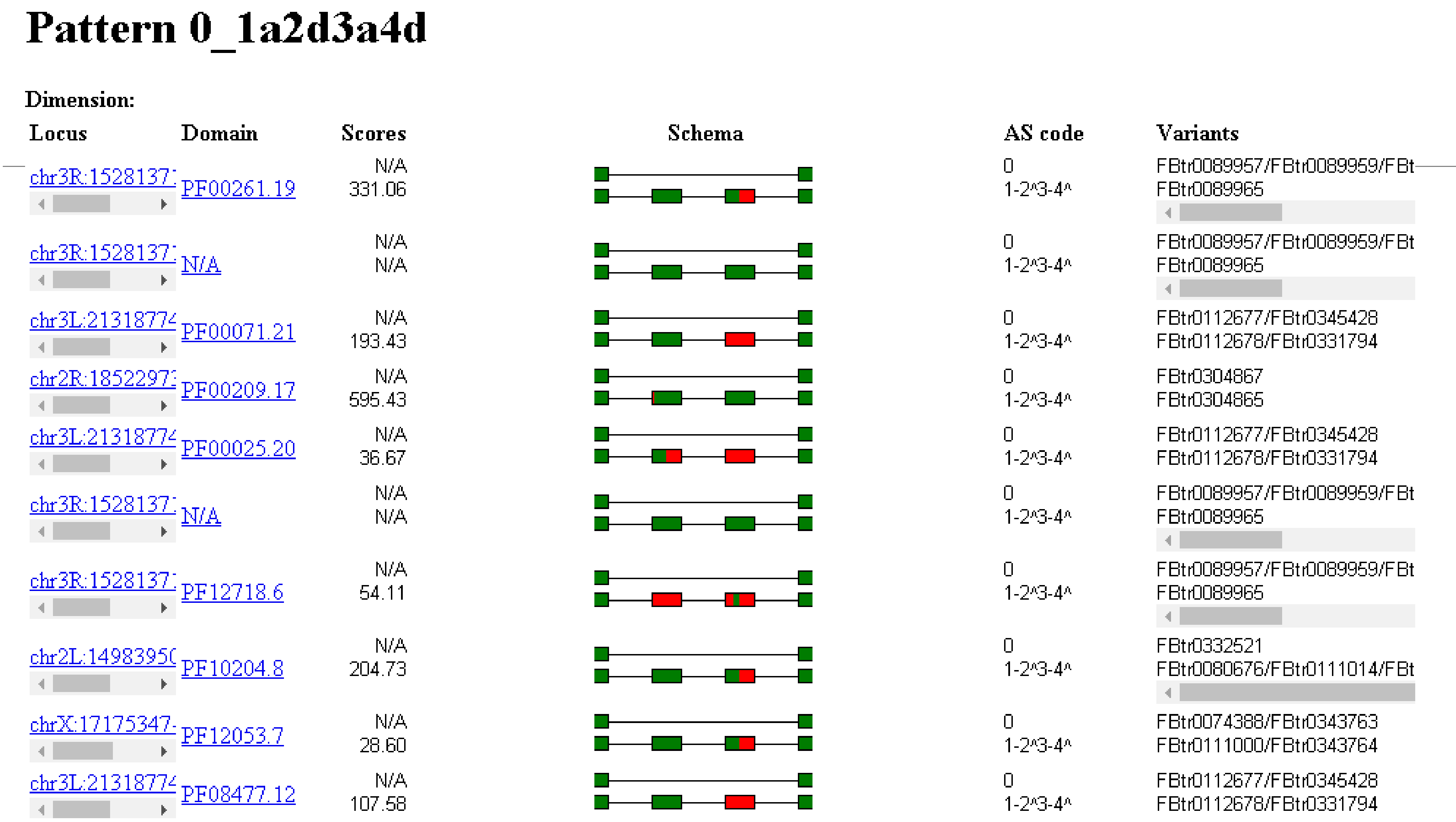
Examples



Pattern 0_1d2a

Dimension:

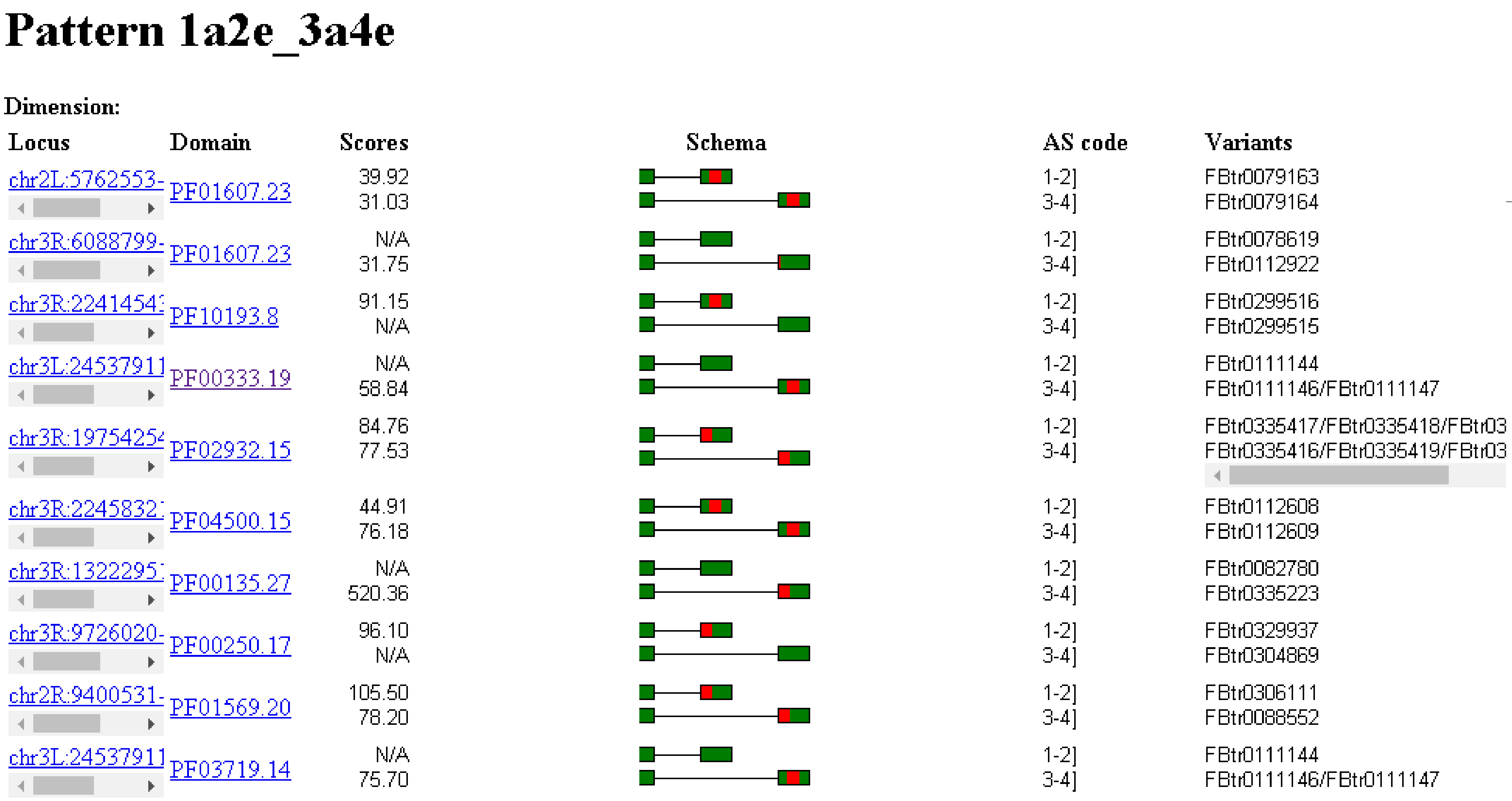
Locus	Domain	Scores	Schema	AS code	Variants
chr3L:13477541	PF02179.15	72.51 72.16		0 1^2-	FBtr0075808/FBtr0113425 FBtr0075806/FBtr0075807
chr3R:2260306	PF01369.19	116.92 117.47		0 1^2-	FBtr0113396/FBtr0334559/FBt FBtr0113395/FBtr0302187
chrX:2344559-2	PF00238.18	43.84 43.84		0 1^2-	FBtr0070405 FBtr0333730
chr3L:9728149-	PF08445.9	22.50 N/A		0 1^2-	FBtr0089411 FBtr0089410/FBtr0114566
chr3R:3002284	PF00615.18	N/A 68.82		0 1^2-	FBtr0300600 FBtr0085553/FBtr0085554/FBt
chrX:7933774-7	PF00411.18	164.54 164.54		0 1^2-	FBtr0071095 FBtr0071094
chrX:1471348-1	PF13903.5	N/A 76.91		0 1^2-	FBtr0100198 FBtr0070255/FBtr0070256
chr2R:2242625	PF01529.19	76.68 64.88		0 1^2-	FBtr0071841 FBtr0071840
chr3L:1608067-	PF16870.4	148.32 183.83		0 1^2-	FBtr0091793/FBtr0091795/FBt FBtr0091792/FBtr0091794/FBt



Pattern 0_1a2d3a4d

Dimension:

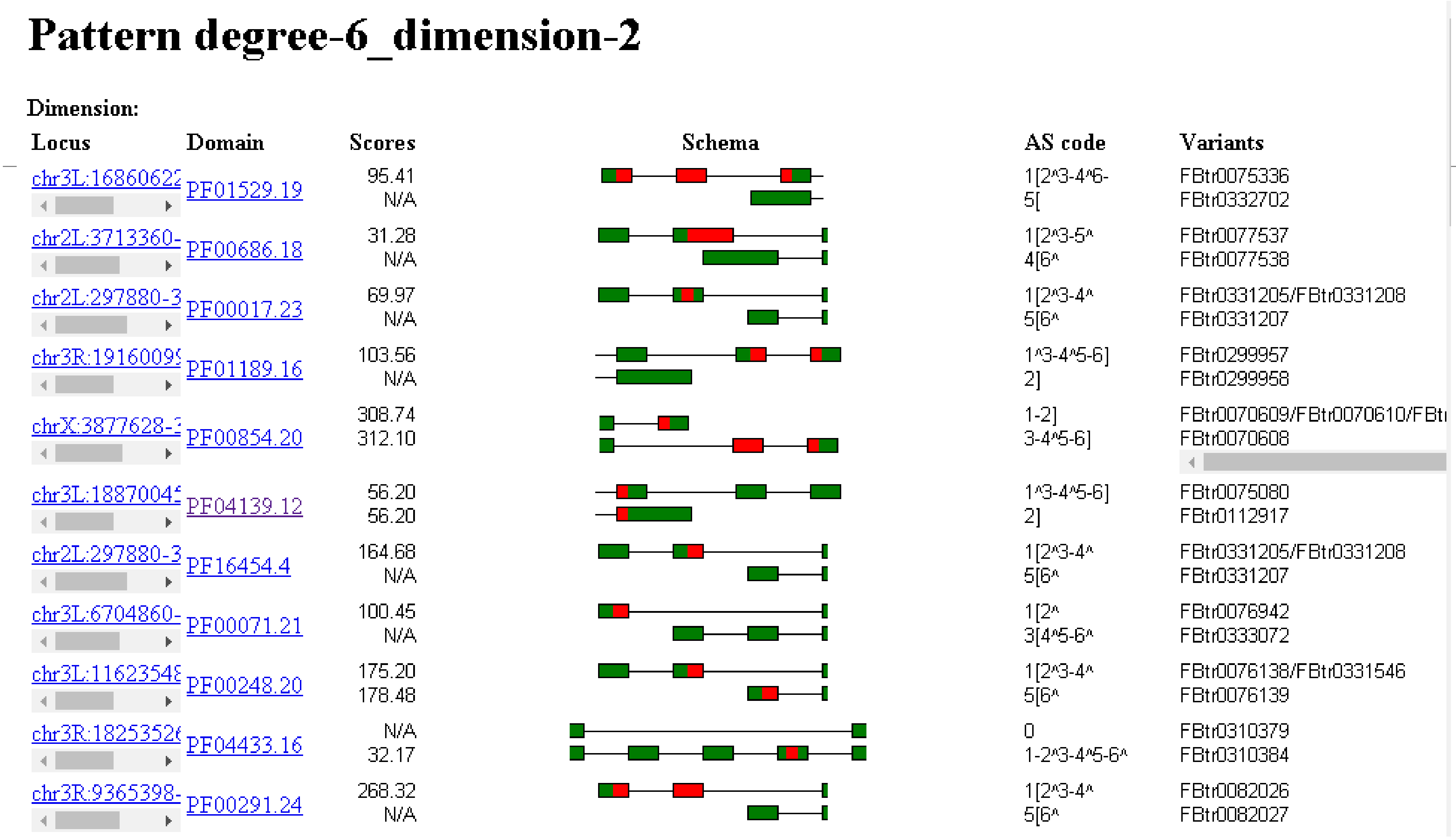
Locus	Domain	Scores	Schema	AS code	Variants
chr3R:1528137	PF00261.19	N/A 331.06		0 1-2^3-4^	FBtr0089957/FBtr0089959/FBt FBtr0089965
chr3R:1528137	N/A	N/A N/A		0 1-2^3-4^	FBtr0089957/FBtr0089959/FBt FBtr0089965
chr3L:21318774	PF00071.21	N/A 193.43		0 1-2^3-4^	FBtr0112677/FBtr0345428 FBtr0112678/FBtr0331794
chr2R:1852297	PF00209.17	N/A 595.43		0 1-2^3-4^	FBtr0304867 FBtr0304865
chr3L:21318774	PF00025.20	N/A 36.67		0 1-2^3-4^	FBtr0112677/FBtr0345428 FBtr0112678/FBtr0331794
chr3R:1528137	N/A	N/A N/A		0 1-2^3-4^	FBtr0089957/FBtr0089959/FBt FBtr0089965
chr3R:1528137	PF12718.6	N/A 54.11		0 1-2^3-4^	FBtr0089957/FBtr0089959/FBt FBtr0089965
chr2L:1498395	PF10204.8	N/A 204.73		0 1-2^3-4^	FBtr0332521 FBtr0080676/FBtr0111014/FBt
chrX:17175347-	PF12053.7	N/A 28.60		0 1-2^3-4^	FBtr0074388/FBtr0343763 FBtr0111000/FBtr0343764
chr3L:21318774	PF08477.12	N/A 107.58		0 1-2^3-4^	FBtr0112677/FBtr0345428 FBtr0112678/FBtr0331794



Pattern 1a2e_3a4e

Dimension:

Locus	Domain	Scores	Schema	AS code	Variants
chr2L:5762553-	PF01607.23	39.92 31.03		1-2] 3-4]	FBtr0079163 FBtr0079164
chr3R:6088799-	PF01607.23	N/A 31.75		1-2] 3-4]	FBtr0078619 FBtr0112922
chr3R:2241454	PF10193.8	91.15 N/A		1-2] 3-4]	FBtr0299516 FBtr0299515
chr3L:24537911	PF00333.19	N/A 58.84		1-2] 3-4]	FBtr0111144 FBtr0111146/FBtr0111147
chr3R:1975425	PF02932.15	84.76 77.53		1-2] 3-4]	FBtr0335417/FBtr0335418/FBtr03 FBtr0335416/FBtr0335419/FBtr03
chr3R:2245832	PF04500.15	44.91 76.18		1-2] 3-4]	FBtr0112608 FBtr0112609
chr3R:1322295	PF00135.27	N/A 520.36		1-2] 3-4]	FBtr0082780 FBtr0335223
chr3R:9726020-	PF00250.17	96.10 N/A		1-2] 3-4]	FBtr0329937 FBtr0304869
chr2R:9400531-	PF01569.20	105.50 78.20		1-2] 3-4]	FBtr0306111 FBtr0088552
chr3L:24537911	PF03719.14	N/A 75.70		1-2] 3-4]	FBtr0111144 FBtr0111146/FBtr0111147

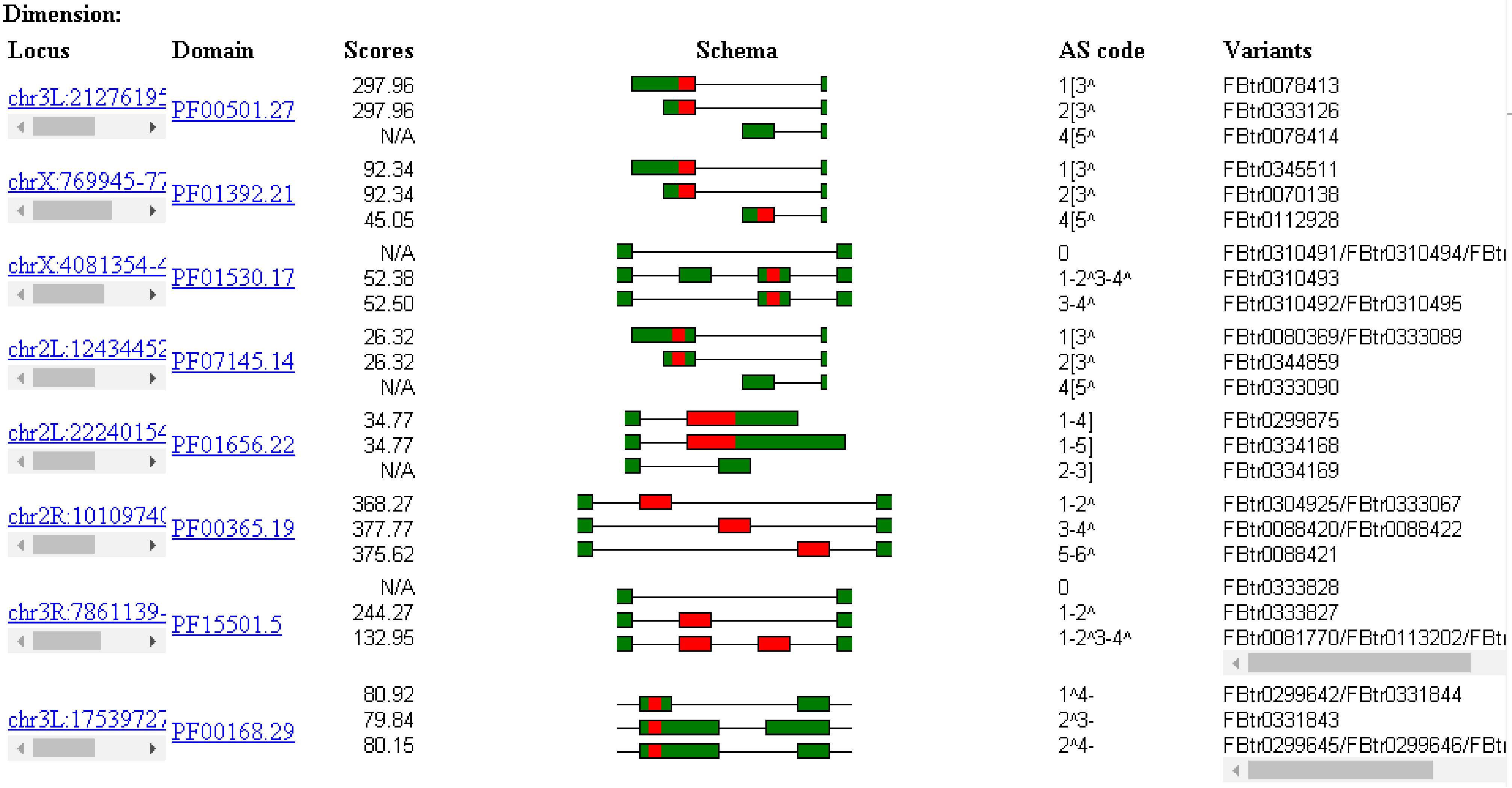


Pattern degree-6_dimension-2

Dimension:

Locus	Domain	Scores	Schema	AS code	Variants
chr3L:16860622	PF01529.19	95.41 N/A		1[2^3-4^6- 5[	FBtr0075336 FBtr0332702
chr2L:3713360-	PF00686.18	31.28 N/A		1[2^3-5^ 4[6^	FBtr0077537 FBtr0077538
chr2L:297880-3	PF00017.23	69.97 N/A		1[2^3-4^ 5[6^	FBtr0331205/FBtr0331208 FBtr0331207
chr3R:19160099	PF01189.16	103.56 N/A		1^3-4^5-6] 2]	FBtr0299957 FBtr0299958
chrX:3877628-3	PF00854.20	308.74 312.10		1-2] 3-4^5-6]	FBtr0070609/FBtr0070610/FBtı FBtr0070608
chr3L:18870045	PF04139.12	56.20 56.20		1^3-4^5-6] 2]	FBtr0075080 FBtr0112917
chr2L:297880-3	PF16454.4	164.68 N/A		1[2^3-4^ 5[6^	FBtr0331205/FBtr0331208 FBtr0331207
chr3L:6704860-	PF00071.21	100.45 N/A		1[2^ 3[4^5-6^	FBtr0076942 FBtr0333072
chr3L:11623548	PF00248.20	175.20 178.48		1[2^3-4^ 5[6^	FBtr0076138/FBtr0331546 FBtr0076139
chr3R:18253526	PF04433.16	N/A 32.17		0 1-2^3-4^5-6^	FBtr0310379 FBtr0310384
chr3R:9365398-	PF00291.24	268.32 N/A		1[2^3-4^ 5[6^	FBtr0082026 FBtr0082027

Pattern degree-6_dimension-3



Dimension:

Locus	Domain	Scores	Schema	AS code	Variants
chr3L:2127619	PF00501.27	297.96		1[3^	FBtr0078413
		297.96		2[3^	FBtr0333126
		N/A		4[5^	FBtr0078414
chrX:769945-77	PF01392.21	92.34		1[3^	FBtr0345511
		92.34		2[3^	FBtr0070138
		45.05		4[5^	FBtr0112928
chrX:4081354-	PF01530.17	N/A		0	FBtr0310491/FBtr0310494/FBt
		52.38		1-2^3-4^	FBtr0310493
		52.50		3-4^	FBtr0310492/FBtr0310495
chr2L:1243445	PF07145.14	26.32		1[3^	FBtr0080369/FBtr0333089
		26.32		2[3^	FBtr0344859
		N/A		4[5^	FBtr0333090
chr2L:2224015	PF01656.22	34.77		1-4]	FBtr0299875
		34.77		1-5]	FBtr0334168
		N/A		2-3]	FBtr0334169
chr2R:1010974	PF00365.19	368.27		1-2^	FBtr0304925/FBtr0333067
		377.77		3-4^	FBtr0088420/FBtr0088422
		375.62		5-6^	FBtr0088421
chr3R:7861139-	PF15501.5	N/A		0	FBtr0333828
		244.27		1-2^	FBtr0333827
		132.95		1-2^3-4^	FBtr0081770/FBtr0113202/FBt
chr3L:1753972	PF00168.29	80.92		1^4-	FBtr0299642/FBtr0331844
		79.84		2^3-	FBtr0331843
		80.15		2^4-	FBtr0299645/FBtr0299646/FBt

AstaFunk

http://astafunk.sammeth.net

Vitor Lima Coelho & Michael Sammeth

micha@sammeth.net

vitorcoelho@biof.ufrj.br

CNPq
elho Nacional de Desenvolvimento
ífico e Tecnológico