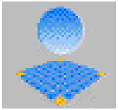
Evolutionary process of a tetranucleotide microsatellite locus in Acipenseriforms

Eric Rivals

LIRMM - Méthodes Algorithmes pour la Bioinfo

`www.lirmm.fr/~rivals`

Laboratoire
d'Informatique
de Robotique
et de Microélectronique
de Montpellier



Paddlefishes and Sturgeons

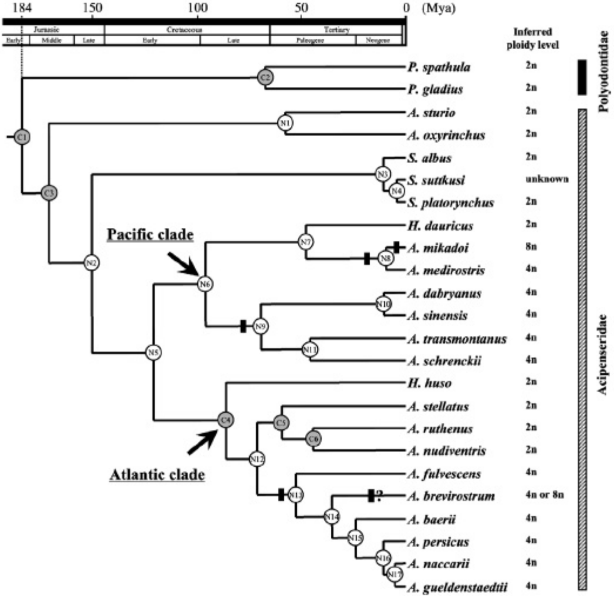
© Ken Lucas / www.ardea.com

Paddlefishes and Sturgeons

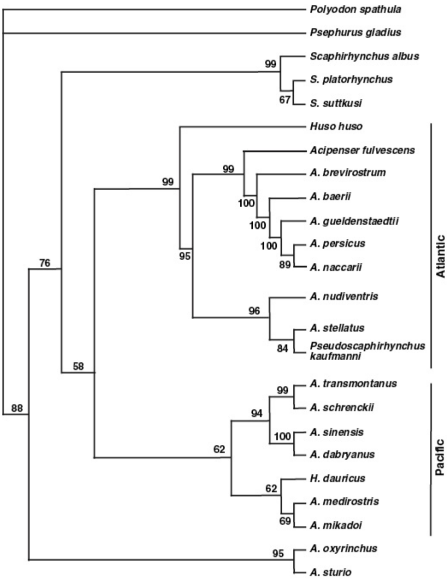
Characteristics

- Acipenseriforms : Two subfamilies, 4 genera, 27 extant species
- 200 million years of evolution ; living fossils
- Many species produce viable hybrids
- Several levels of polyploidy

Phylogeny from [Peng et al. 06]



Phylogeny from [Krieger et al. 08]



A composite microsatellite Spl-106

Species	Lg	Repeat like	Repeat
AG	199	$(taaaa)(aaaa)(taaa)$	$(taga)^2(taaa)^{12}$
AT	198	$(taaa-)(aaaa)(taaa)$	$(taga)^2(taaa)(taga)^2(taaa)^7(gaaa)^2$
AS	170	–	$(taaa)^9(gaaa)^4$
HD	263	$(taaa)(aaaa)(taaa)$	$(taga)^{23}(tagga)((taaa)(aaaa))^2(taaa)^2$

94 orthologous alleles ; lengths comprise in $[139, 263]$ bps ; size homoplasy

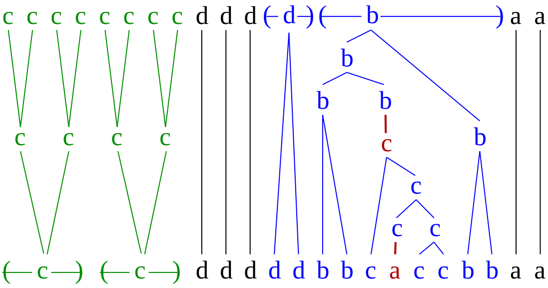
Microsatellite sequence recoding

Encoding each variant repeat unit by a distinct symbol

	Repeat like			Repeat				
Sequence	$(taaaa)$	$(aaaa)$	$(taaa)$	$(taga)^2$	$(taaa)^{12}$			
Map	O'	J	O	U^2	O^{12}			
Sequence	$(taaa-)$	$(aaaa)$	$(taaa)$	$(taga)^2$	$(taaa)$	$(taga)^2$	$(taaa)^7$	$(gaaa)^2$
Map	O	J	O	U^2	O	U^2	O^7	Q^2
Sequence	–	–	–	$(taaa)^9$	$(gaaa)^4$			
Map	–	–	–	O^9	Q^4			

Allele comparison by alignment

Example of an alignment of 2 maps



- contexte dependency - non commutativity
- Version 1 : correspondance with a graph algorithm
  stable max in a *circle graph* [Bérard Rivals *J Comp Bio* 2003]
- Version 2 : dynamic programming
  [Bérard et al., *Evol. Bioinfo.* 2006]

Protocol of the analysis

Input : set of sequences (maps)

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]
2. Inference of evolutionary tree from the distances

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]
2. Inference of evolutionary tree from the distances
   FastME [Desper, Gascuel, 2002]

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]
2. Inference of evolutionary tree from the distances
   FastME [Desper, Gascuel, 2002]
3. Test robustness of the trees w.r.t. alignment parameters

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]
2. Inference of evolutionary tree from the distances
   FastME [Desper, Gascuel, 2002]
3. Test robustness of the trees w.r.t. alignment parameters
   Criterion : Percentage of explained variance (VAF)

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]
2. Inference of evolutionary tree from the distances
   FastME [Desper, Gascuel, 2002]
3. Test robustness of the trees w.r.t. alignment parameters
   Criterion : Percentage of explained variance (VAF)
4. Assess confidence of the tree internal nodes

Protocol of the analysis

Input : set of sequences (maps)

1. Comparison all against all ⇒ pairwise distance matrix
   MS_Align [Bérard, Rivals, 2003]
2. Inference of evolutionary tree from the distances
   FastME [Desper, Gascuel, 2002]
3. Test robustness of the trees w.r.t. alignment parameters
   Criterion : Percentage of explained variance (VAF)
4. Assess confidence of the tree internal nodes
   Criterion : Rate of elementary well designed quartets (Re)
   Qualitree [Garreta, Guénoche, 2000]

Rate of elementary well designed quartets (Re)

For an internal edge e, for all quartets (i, j, k, l)
s.t. e splits (i, j) and (k, l) :

$R(e)$ = percentage of these quartets satisfying

$$D(i,j) + D(k,l) < \min\left(D(i,l) + D(j,k), D(i,k) + D(j,l)\right)$$

where
$D(i,j)$: alignment distance between i and j

Value in $[0, 1]$ [Garreta, Guénoche, 2000]

Evolution of Spl-106

Development of the composite microsatellite :

- from $(taga)^m$ in *S. platorynchus*
- to $(taga)^m(taaa)^n$ with $m > n$
- to $(taga)^m(taaa)^n$ with $n > m$ in most Atlantic species
- to $(taaa)^p(gaaa)^q$ in *A. sinensis, A. dabryanus*, most recent state

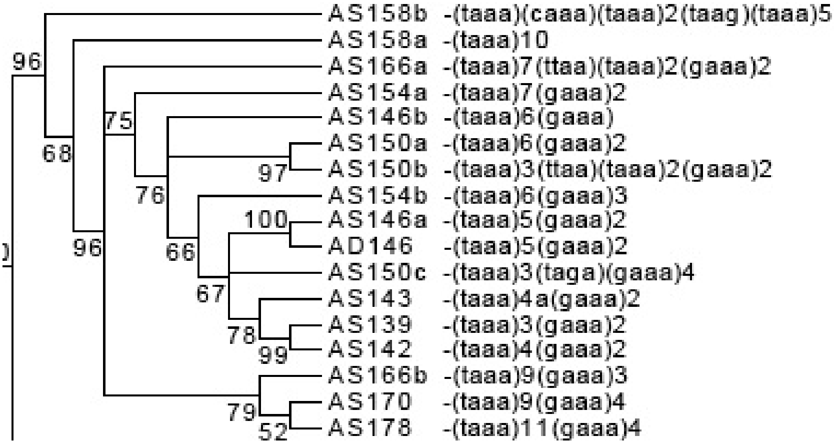
Evidence for sturgeons phylogeny

Spl-106 and flanking regions phylogenies support :

- *Huso* species are not monophyletic
- Uniformity of type of repeat region in Atlantic clade
- Separate evolution of Atlantic and Pacific clades
- An early separation of *A. sturio* and *A. oxyrinchus* from other subfamilies

Spl-106 allows to investigate intra-species relationships

Phylogeny of *A. sinensis* with Spl-106



Co-authors :

- ZhaoJun Shao, Patrick Berrebi, ISEM, Montpellier.
- Bin Zhu, Na Zhao and Jianbo Chang, Wuhan, China
- Sovan Lek, Ichthyology & Ecological Modelling, UMR 5174 Univ. Paul Sabatier, Toulouse

Thanks for your attention

Evolutionary model

- Substitution : WGY → WKY
- Deletion : WGY → WY
- Insertion (dual) : WY → WGY
- Tandem duplication : WKY → WKKY
- Tandem contraction (dual) : WKKY → WKY

Evolutionary model

- Substitution : WGY → WKY
- Deletion : WGY → WY
- Insertion (dual) : WY → WGY
- Tandem duplication : WKY → WKKY
- Tandem contraction (dual) : WKKY → WKY

⇒ variation in their number of units

Percentage of explained variance (VAF)

$$\mathrm{VAF} = 1 - \frac{\sum_{(i,j):i<j}(D(i,j) - T(i,j))^2}{\sum_{(i,j):i<j}(D(i,j) - D_m)^2}$$

where

$D(i,j)$:	alignment distance between i and j
$T(i,j)$:	tree distance between i and j
D_m :	average alignment distance over all pairs (i,j)

Value in $[0,1]$