

# **Analisando e Editando Seqüências e Alinhamentos Moleculares com Python.**

**Frederico Gonzalez Colombo Arnoldi**

UNESP – Rio Claro - Departamento de Biologia Celular e Molecular

- Apresentação.
- Como o meu trabalho, de um biólogo, poderia ser mais interessante aqui?
- Expectativa.

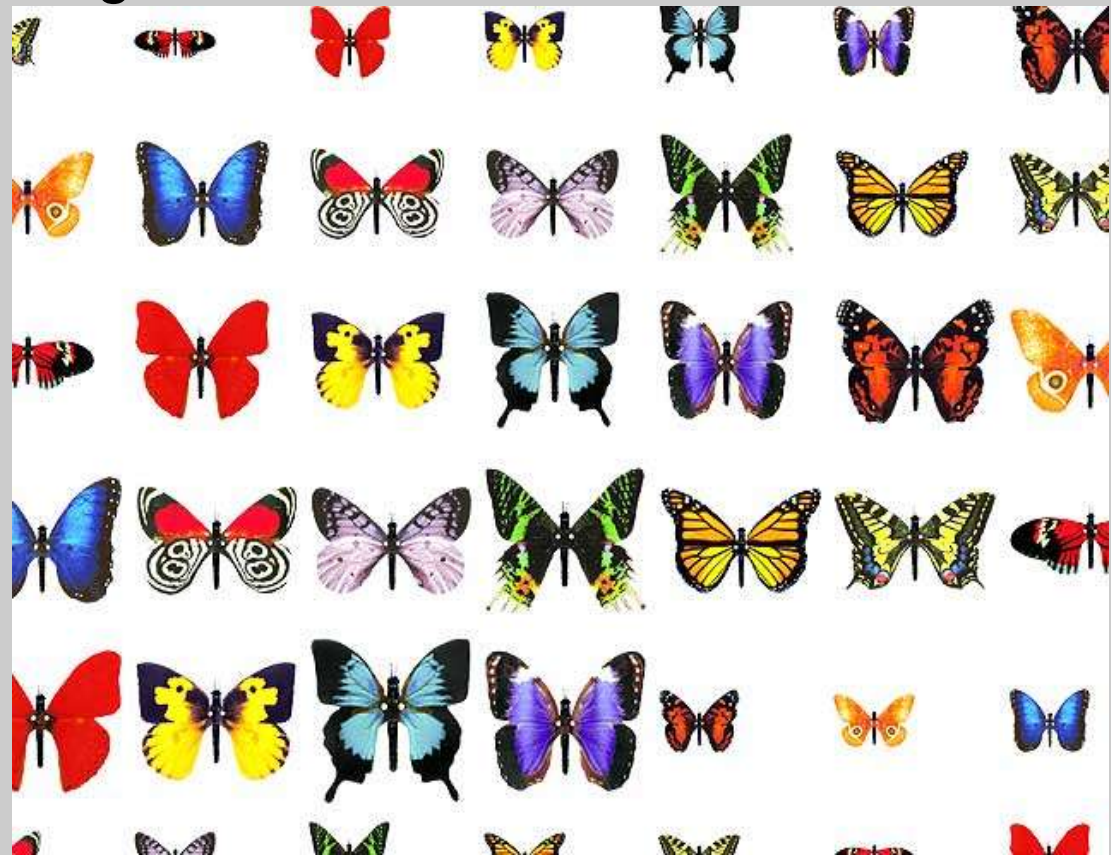
# O PROBLEMA BIOLÓGICO

- Biologia Comparada = Comparar organismos ou parte deles.

Exemplos: Indivíduo Sadio/Doente,

Microorganismo resistente/suscetível

Algumas comparações  
são intuitivamente fáceis.



- Mas e....



- Primeiro passo: Estabelecer homologias.

Estruturas homólogas são aquelas que apresentam uma mesma origem evolutiva.

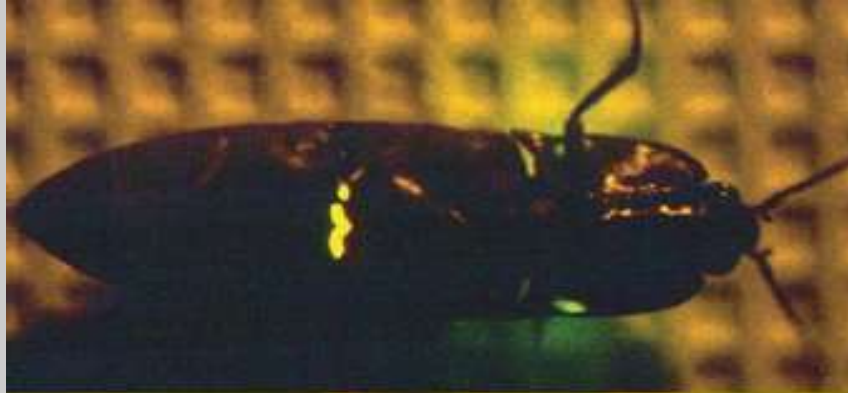
## **Objetivos do meu projeto:**

Identificar, nas luciferases de coleópteros, os resíduos responsáveis pela modulação da cor da luminescência por elas emitida.

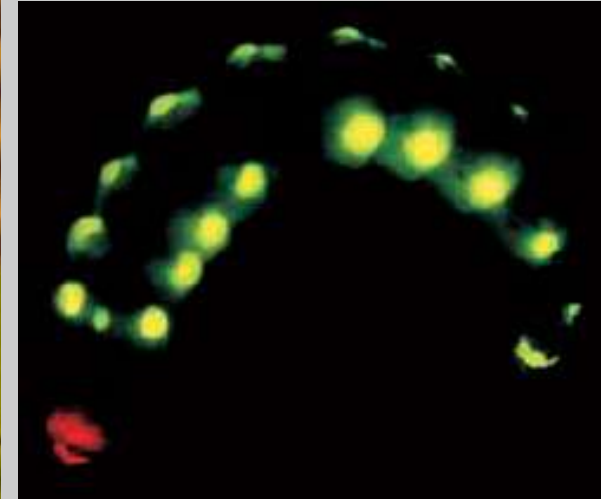
As três maiores famílias de coleópteros  
luminescentes.



Lampyridae



Elateridae

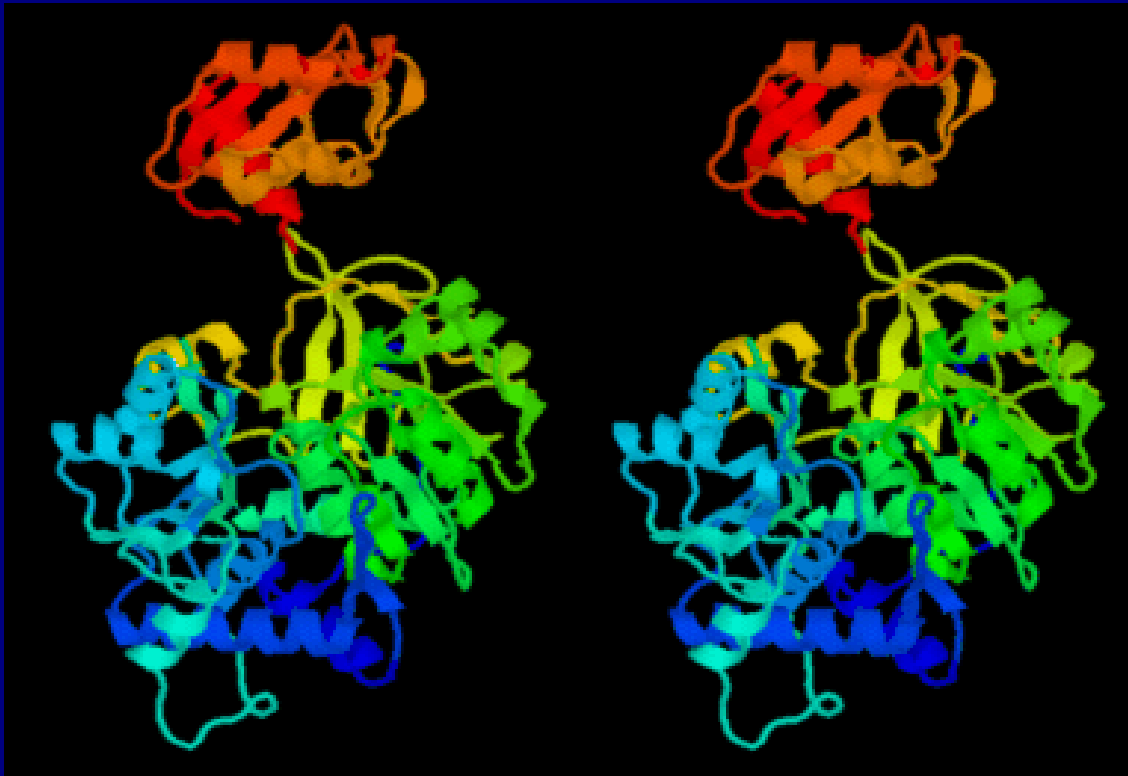


Phengodidae

O sistema que gera luz em coleópteros é formado basicamente por: um substrato denominado de **luciferina**, uma enzima denominada **luciferase**, **ATP** e **Oxigênio**.

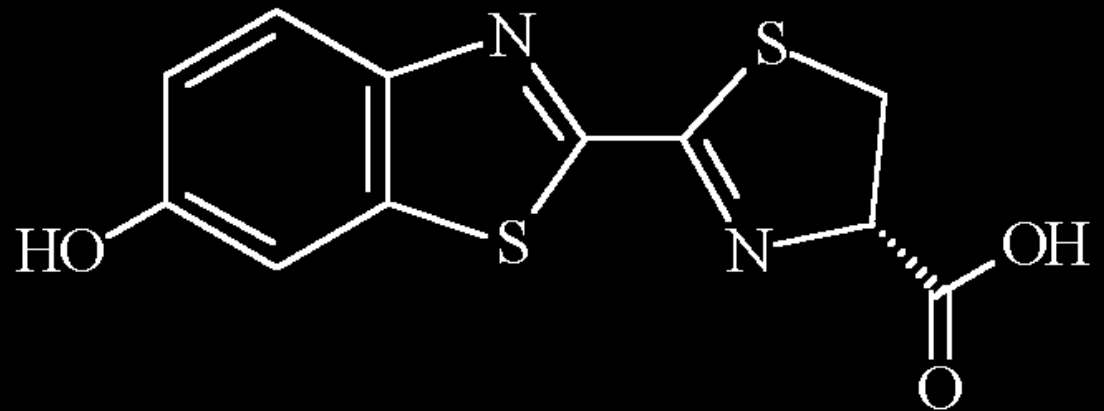
*luciferin*

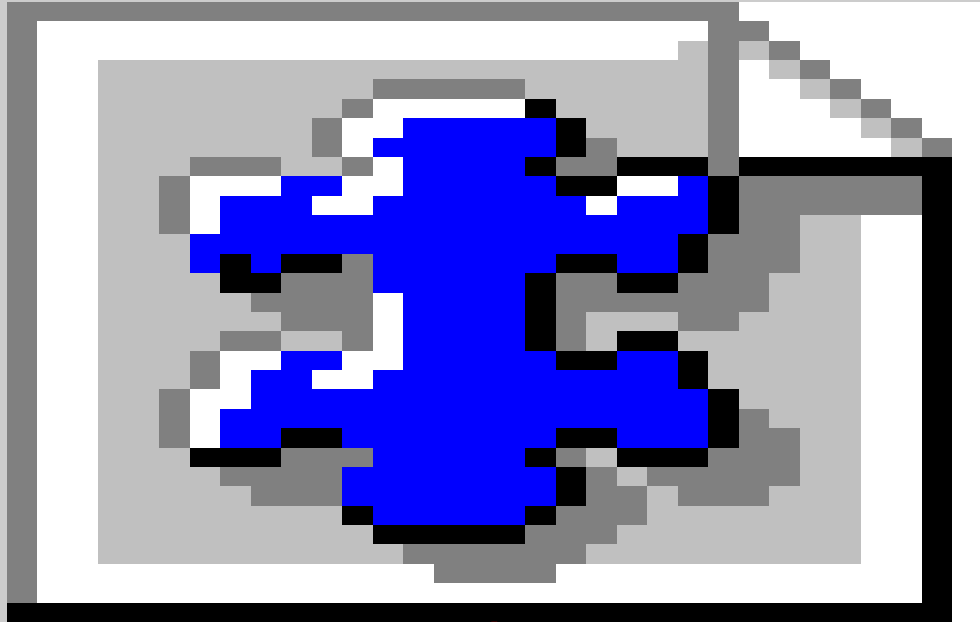
*luciferase*



Modelo da  
luciferase de  
Coleóptera

Luciferina de  
Coleóptera





Existem softwares já desenvolvidos para o propósito?

- Sim.

Desenvolver outro?

Programas poderosos / áridos.

Programas amigáveis / sem recursos.

Programas poderosos e amigáveis / não disponíveis para todos os Sistemas Operacionais.

**MPAling**

# Trabalhando interativamente com o MPAlign:

```
Ruler      .....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.
           .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....
P_pen      ---MSIENNILIGPPPYPLEEGTAGEQLHRAISRYAAVPGTLAYTDVHTELEVITYKEFLDVTCLAEAMKNYGLGLQHTISVCSENCVQFFMPICAALYVGV
P_miy      --MEDDSKHIMHGHRHSILWEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYFEMSCRLAETMKRYGLGLQHIIAVCSETSLQFFMPVCGALFIGV
Macro      --MED-EKNIIHGPEPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDAHIEVNITYAEYFEMSCRLAEAMXRYGLGLKHRIIVVCEENSLQFFMPVLGALFIGV
C_dis      --MEE-DKNIMYGPAPFSPLEEGTAGEQLHKAMKRYAQIPGTIAFTA AHVEVNVTYAEYFEMACRLAETMKRYGLGLDHRIAVCSEENSLQFFMPVCGALFIGV
P_pyr      --MED-AKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDAHIEVNITYAEYFEMSVRLAEAMKRYGLNTNHRIVVCEENSLQFFMPVLGALFIGV
P_rufa     --MEDDSKHIMHGHRHSILWEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYFEMSCRLAETMKRYGLGLQHIIAVCSEENSLQFFMPVCGALFIGV
l_noc      --MED-AKNIMHGAPAPFYPLEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYFEMACRLAETMKRYGLGLQHIIAVCSEENSLQFFMPVCGALFIGV
H_par      -MEMEKEENVVYGPLPFYPIEEGSAGIQLHKYMQYAKL -GAI AFSNALTGVDISYQEYFDITCRLAEAMKNYGMKQEGTIALCSENCEEFFIPVLAGLYIGV
L_min      -MEMEKEENVVYGPLPFYPIEEGSAGIQLHKYMHQYAKL -GAI AFSNALTGVDISYQEYFDITCRLAEAMKNFGMKPEEHIALCSENCEEFFIPVLAGLYIGV
L_cru      MENMENDENIVVGPKPFYPIEEGSAGTQLRKYMERYAKL -GAI AFTNAVTVGDYSYAEYLEKSCCLGKALQNYGLVVDGRIALCSENCEEFFIPVLAGLYIGV
L_lat      MENMENDENIVYGPEPFYPIEEGSAGAQLRKYMDRYAKL -GAI AFTNALTGV DYTAEYLEKSCCLGEALKNYGLVVDGRIALCSENCEEFFIPVLAGLYIGV
```

# Alinhando as seqüências com MALIGN:

**Alignment with Malign** - Parameters Files Malign Manual

**Parameter of Malign**

Gap costs: A C C G T G A G

Leading: 2 Internal: 3 Trailing: 2 Extra: 2

Complex Costs:

A	C	G	T	
0	1	1	1	A
1	0	1	1	C
1	1	0	1	G
1	1	1	0	T

Tv:Ts costs

Output format: paup

Reporting options: complete

Alignment construction: Procedures: build

Branch Swapping: aspr

Alignment Cost Function: Score: 2 Cladogram Based Costs: keeptrees 100

treerandorde 1 Branch Swapping: spr

Parameter file: parameter.txt

Others commands:  + Add command

Save

Close

**Alignment with Malign** - Parameters Files Malign Manual

**First Alignment**

Input File: temp.mal Output File: Group File:

Second Alignment

**Second Alignment**

Output File: Parameter File 2: Group File 2:

Third Alignment

**Third Alignment**

Output File 3: Parameter File: Group File 3:

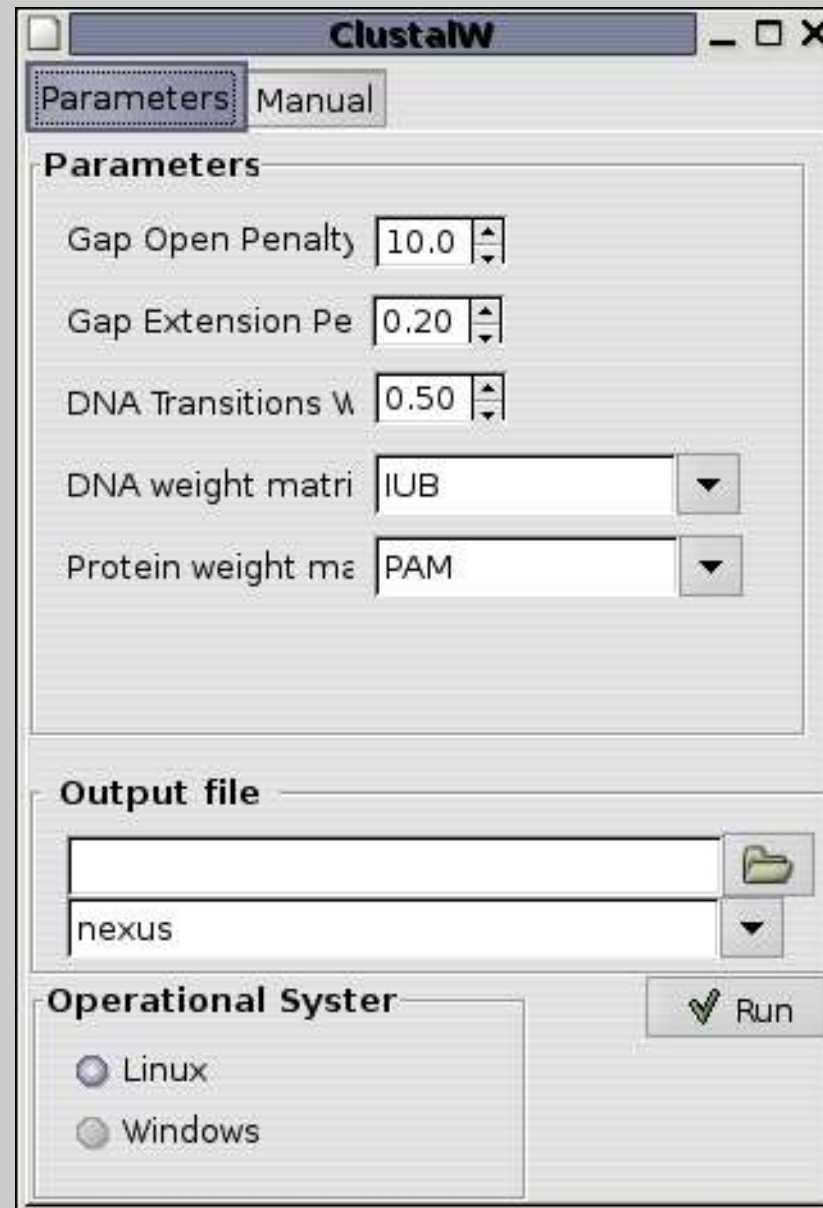
**Operational System**

Linux  Windows

Execute

Close

# Alinhando com CLUSTALW:



The image shows a screenshot of the ClustalW software interface. The window title is "ClustalW". There are two tabs: "Parameters" (selected) and "Manual". The "Parameters" section contains several settings:

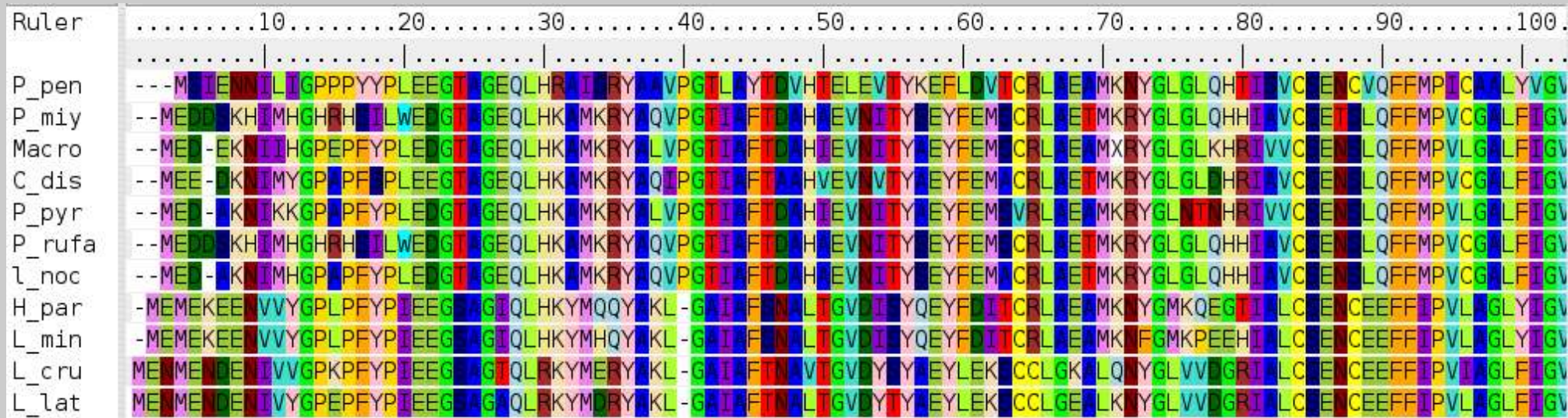
- Gap Open Penalty: 10.0
- Gap Extension Pe: 0.20
- DNA Transitions W: 0.50
- DNA weight matri: IUB
- Protein weight ma: PAM

The "Output file" section has a text box containing "nexus" and a folder icon to its right.

The "Operational System" section has two radio buttons: "Linux" (selected) and "Windows".

A "Run" button with a checkmark icon is located at the bottom right of the dialog.

# Colorindo cada resíduo:


























# Enfatizando a conservação:

Ruler	.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....
P_pen	--MSIENNILIPPPYYPLEETAEEDLHRAISRVAVPTLNYTDVHTELEVTKFLDVTCRAEAMKNYLGLQHTSVSENVCVQIMICAAIYV
P_miy	--MEDDSKHIMHHRHSILWDTAEEDLHKAMKRYAQVPTIIFDAHAIEVNITSYFEMSCR AETMKRYLGLQHHAVSENLSLQIMVCGAFI
Macro	--MED-EKNIHPEPFYPLEDTAEEDLHKAMKRYALVPTIIFDAHIEVNITAYFEMSCR AEAMXRYLGLKHRVVENSLQIMVLGAFI
C_dis	--MEE-DKNIMYPAPFSPLPETAEEDLHKAMKRYAQIPTIIFTAHVEVNVTAYFEMACR AETMKRYLGLDHRAVSENLSLQIMVCGAFI
P_pyr	--MED-AKNIKYPAPFYPLEDTAEEDLHKAMKRYALVPTIIFDAHIEVNITAYFEMSVR AEAMKRYLNTNHRVVENSLQIMVLGAFI
P_rufa	--MEDDSKHIMHHRHSILWDTAEEDLHKAMKRYAQVPTIIFDAHAIEVNITSYFEMSCR AETMKRYLGLQHHAVSENLSLQIMVCGAFI
l_noc	--MED-AKNIMHPAPFYPLEDTAEEDLHKAMKRYAQVPTIIFDAHAIEVNITSYFEMACR AETMKRYLGLQHHAVSENLSLQIMVCGAFI
H_par	-MEMEKEENVVYPLPFYPIEESAGIDLHKYMQQYAKL-AIIFSNALTGVDISQYFDITCR AEAMKNYMKQEGTALSENCEEIIVLAGYI
L_min	-MEMEKEENVVYPLPFYPIEESAGIDLHKYMHQYAKL-AIIFSNALTGVDISQYFDITCR AEAMKNFMKPEEHALSENCEEIIVLAGYI
L_cru	MENMENDENIVVPKPFYPIEESAGIDLRKYMERYAKL-AIIFNAVTGVDYSAYLEKSCCGKALQNYLVVDGRALSENCEEIIVVIAGFI
L_lat	MENMENDENIVYPEPFYPIEESAGIDLRKYMDRYAKL-AIIFTNALTGVDTYAYLEKSCCGEALKNYLVVDGRALSENCEEIIVLAGFI


# Personalizando:


**Set Up** [minimize] [maximize] [close]

**Colors**

Symbol	Color	Complementary	Symbol	Color	Complementary
A		T	-		-
T		A	H		
C		G	Q		
G		C	I		
R		Y	D		
Y		R	V		
W		S	E		
S		W	L		
K		M	P		
M		K	F		
N		N	X		
*		*			

**Categories**

**Conserved** 

**Semi-conserved (75%)** 

**Conserved regions**

Entropy(%)

Minimum length

**Close**

# Trabalhando internamente às seqüências:

The screenshot shows a sequence editor window with a search dialog box open. The search dialog has three sections: ORFs (set to AUG), Restriction Sites (set to SmaI), and Others (set to VI). The main window displays a sequence alignment with a ruler at the top. The search results window on the right lists the following findings:

Sequence	Position
VI on P_miy	Position 148
VI on P_rufa	Position 148
VI on P_rufa	Position 153
VI on P_rufa	Position 244
VI on Macro	Position 148
VI on Macro	Position 153
VI on Macro	Position 266
VI on P_pyr	Position 148
VI on P_pyr	Position 153
VI on P_pyr	Position 244
VI on I_noc	Position 153
VI on I_noc	Position 244

# Criando seqüência consensus:

The image shows a screenshot of a sequence alignment editor window titled "Editor". The window has a menu bar with "Arquivo", "Edit", "Aligning", "Sequence", "Alignment", "Coloring", "SetUp", and "Help". Below the menu bar is a toolbar with icons for file operations and alignment. The main area displays a multiple sequence alignment of 11 protein sequences, with a ruler at the top indicating positions from 10 to 100. The sequences are: P\_miy, P\_rufa, Macro, P\_pyr, l\_noc, C\_dis, H\_par, L\_min, L\_cru, L\_lat, and P\_pen. The consensus sequence is shown at the bottom. The alignment is color-coded, with yellow highlighting conserved residues and blue highlighting gaps or less conserved residues.

Name	Sequence
Ruler	.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.
P_miy	--MEDDSKHIMHHRHSILWDTAGELHKAMKRYQVPITIFDAHAIEVNITSYFEMSCR AETMKRY LGLQHH AVSPTSLQFMVCGA FIV
P_rufa	--MEDDSKHIMHHRHSILWDTAGELHKAMKRYQVPITIFDAHAIEVNITSYFEMSCR AETMKRY LGLQHH AVSPTSLQFMVCGA FIV
Macro	--MED-EKNIHPEPFYPLDITAGELHKAMKRYLVPITIFDAHIEVNITAYFEMSCR AEAMKRY LGLKHR VVSENSLQFMVLGA FIV
P_pyr	--MED-AKNIKKPAPFYPLDITAGELHKAMKRYLVPITIFDAHIEVNITAYFEMSVR AEAMKRY LNTNHR VVSENSLQFMVLGA FIV
l_noc	--MED-AKNIMHPAPFYPLDITAGELHKAMKRYQVPITIFDAHAIEVNITSYFEMACR AETMKRY LGLQHH AVSENSLQFMVCGA FIV
C_dis	--MEE-DKNIMYPAPFSPLDITAGELHKAMKRYQIPITIFTAHVNVITAYFEMACR AETMKRY LGLDHR AVSENSLQFMVCGA FIV
H_par	--MEMEKEENVVYPLPFYPIEISAGI LHKYMQQYKL-AIFSNALTGVDISQYFDITCR AEAMKNY MKQEGT ALSENCEEI VLAG YIV
L_min	--MEMEKEENVVYPLPFYPIEISAGI LHKYMQQYKL-AIFSNALTGVDISQYFDITCR AEAMKNF MKPEEH ALSENCEEI VLAG YIV
L_cru	MENMENDENIVV PKPFYPIEISAGI LHKYMER YKL-AIFTNVAVTGVDYSAYLEKSCG GKALQNY LVVDGR ALSENCEEI VIAG FIV
L_lat	MENMENDENIVV PEPFYPIEISAGI LHKYMDR YKL-AIFTNALTGVDYTAAYLEKSCG GEALKNY LVVDGR ALSENCEEI VLAG FIV
P_pen	--MSIENNILIPPPYYPLEITAGE LHRASRYAVPTLYTDVHTELEVTKFLDVTCR AEAMKNY LGLQHT SVSNCVQFMICAA YIV
Consensu	.....G.....E G AG QL.....YA.....G A.....Y E.....L.....G.....I CSE.....FF P.....L GV

Gerando Relatórios:

Gerando Relatórios:



## Outros....

### DNA

- Sequência reverso.
- Sequência reverso complemento.

Salva seqüências em formato FASTA.

Salva alinhamento em formato NEXUS.

## Agradecimentos:

Daniel Henrique Debonzi – USP – São Carlos