Université de Lille

# Multiple alignment

Adapted from the courses of the Bonsai team,

CRIStAL UMR 9189

Sylvain.legrand@univ-lille.fr

# Introduction

# Definitions

- Input: *K* sequences

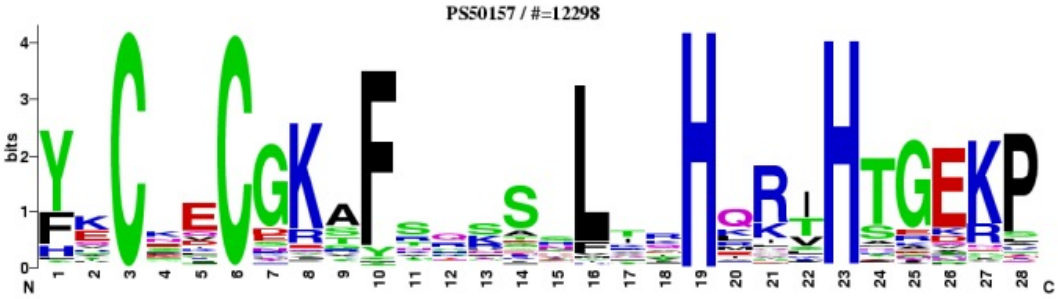
```
C A T G C G A G T A G T A G
C A T G G T A G T A G
C C T G G A G T A C G T A G
C A T G A G C G T A G
```

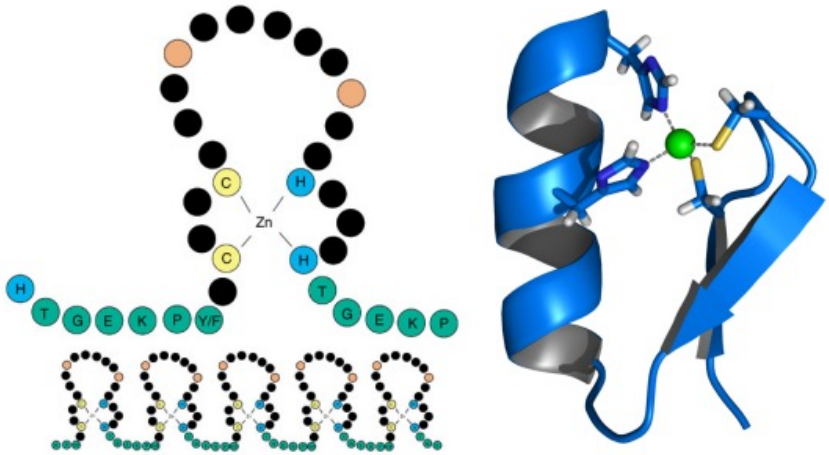
- Output: a table containing the *k* sequences, with indels

```
C A T G C G A G T A - G T A G
C A T G - - - G T A - G T A G
C C T G - G A G T A C G T A G
C A T G - - A G - - C G T A G
```

# Zinc finger pattern (C2H2-type)

```
TYY1_HUMAN YVCPFDGCNKKFAQSTNLKSHILT--H
YKQ8_CAEEL YKCT--VCRKDISSSESLRTHMFKQHH
BASO_HUMAN FQCD--ICKKTFKNACSVKIHHKN-MH
ZG2-9_XENL FVCT--VCGKTYKYKHGLNTHLHS--H
P43_XENBO  LKCSVPGCKRSFRKKRALRIHVSE--H
IKAR_MOUSE FECN--MCGYHSQDRYEFSSHITRGEH
TRA1_CAEEL YKCEFADCEKAFSNASDRAKHQNR-TH
ZN10_HUMAN YKCN--QCGIIFSQNSPFIVHQIA--H
XFIN_XENLA FRCS--ECSRSFTHNSDLTAHMRK--H
TF3A_BUFAM CKCETENCNLAFTTASNMRLHFKR-AH
ZG58_XENLA FVCT--ECNLSFAGLANLRSHQHL--H
P43_XENBO  YRCSYEDCQTVSPTWTALQTHLKK--H
TSH_DROME  FRCV--WCKQSFPTLEALTTHMKDSKH
ZN76_HUMAN FRCGYKGCGRLYTTAHHLKVHERA--H
TF3A_BUFAM YRCPRENCDRTYTTKFNLKSHILT-FH
SUHW_DROAN YACK--ICGKDFTRSYHLKRHQKYSSC
ZN76_HUMAN YTCPEPHCGRGFTSATNYKNHVRI--H
SRYC_DROME FKCN--YCPRDFTNFPNWLKHTRR-RH
EVI1_HUMAN YRCK--YCDRSFSISSNLQRHVRN-IH
```



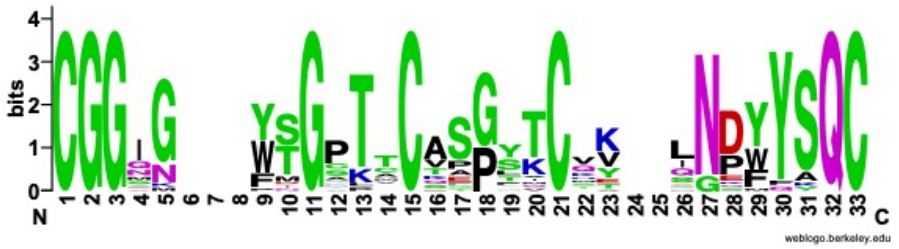

modélisation : motif Prosite

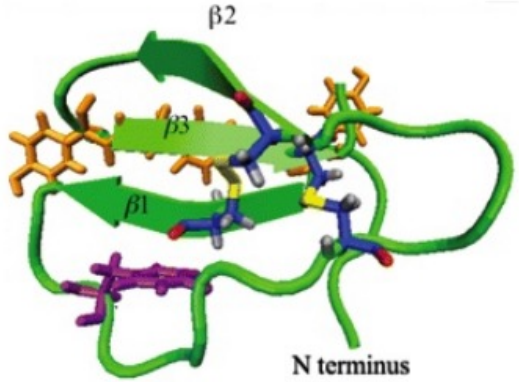
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

# Cellulose binding site

```
HWGQCGGI---GYSGCKTCTSGTTCQYSNDYYSQCL
HYGQCGGI---GYSGPTVCASGTTCQVLNPYYSQCL
QWGQCGGI---GYTGSTTCASPYTCHVLNPYYSQCY
VWGQCGGQ---NWSGPTCCASGSTCVYSNDYYSQCL
LYGQCGGA---GWTGPTTCQAPGTCKVQNQWYSQCL
IWGQCGGN---GWTGATTCASGLKCEKINDWYYQCV
VWGQCGGN---GWTGPTTCASGSTCVKQNDFYSQCL
DWAQCGGN---GWTGPTTCVSPYTCTKQNDWYSQCL
QWGQCGGQ---NYSGPTTCKSPFTCKKINDFYSQCQ
RWQQCGGI---GFTGPTQCEEPYICTKLNDWYSQCL
HWAQCGGI---GFSGPTTCPEPYTCAKDHDIYSQCV
LYEQCGGI---GFDGVTCCSEGLMCMKMGPYYSQCR
VWAQCGGQ---NWSGTPCCTSGNKCVKLNDFYSQCQ
PYGQCGGM---NYSGKTMCSPGFKCVELNEFFSQCD
AYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCV
EYAACGGE---MFMGAKCCKFGLVCYETSGKWSQCR
```

extrait de Prosite, entrée PS00562



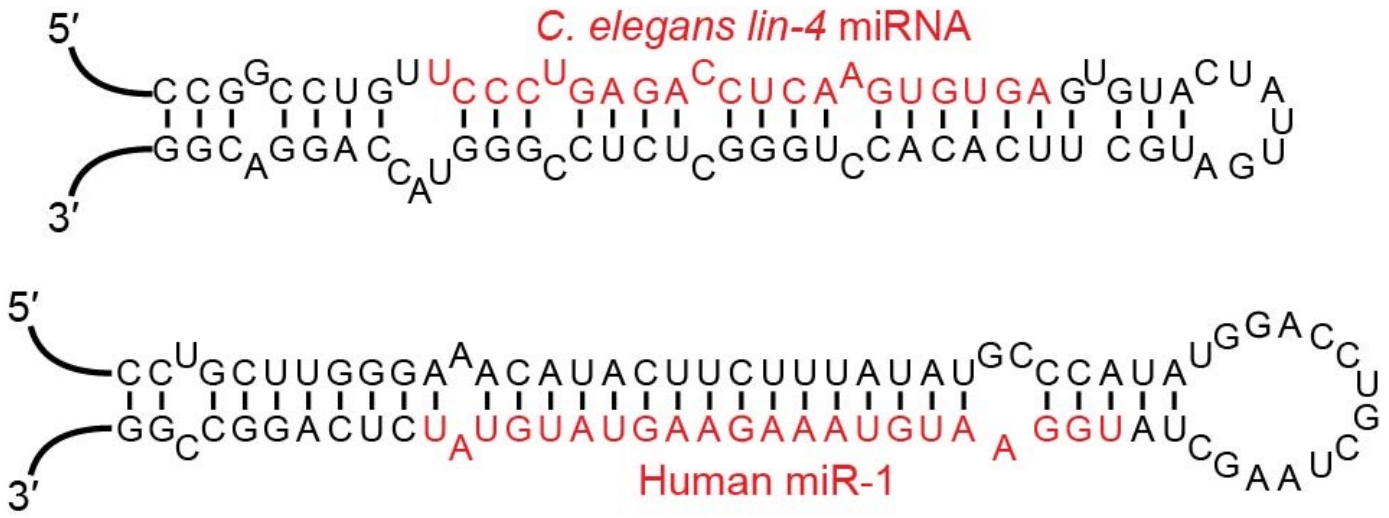

C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C


the 4 cysteines are involved in disulfide bonds (SS-bond)

# RNA structure

- We have a family of RNAs possessing the same secondary structure

- For a given structure pairing:
  - if a base mutates in the RNA structure, the base that matches it must mutate too… → **compensatory mutation**



- Example

```
G  A  G  C  C  C  A  G  U  U  C
   A  G  G  A  C  U  C  U  U  C
A  A  U  C  A  C  C  C  G  A  U
```

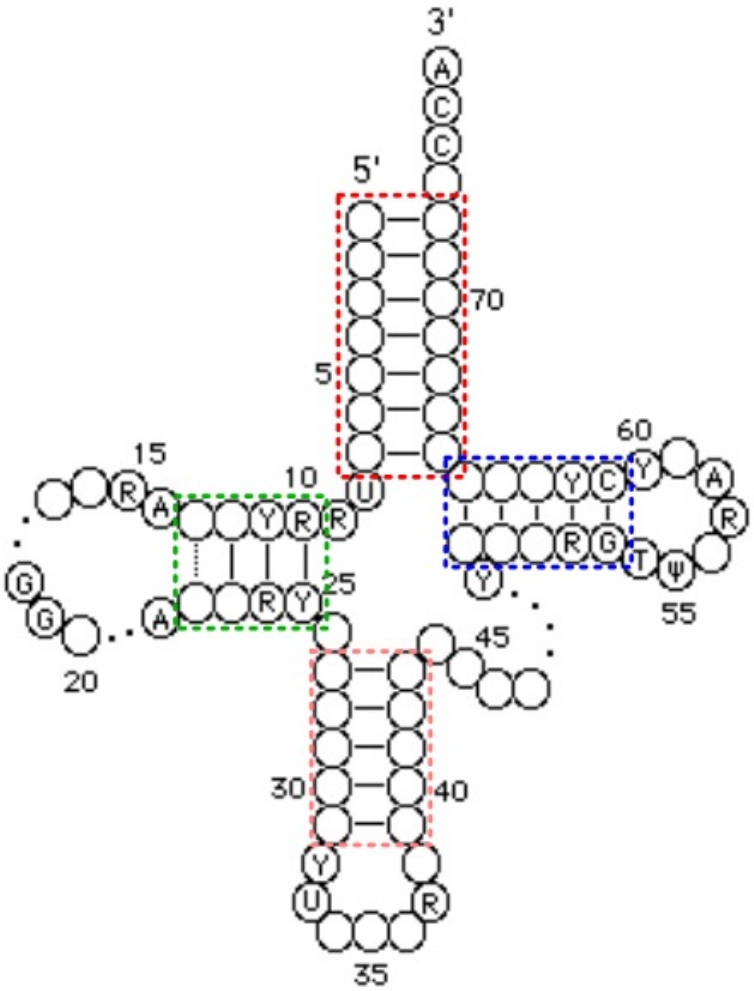
- Step 1: **multiple alignment** construction

```
G  A  G  C  -  C  C  A  G  U  U  C
-  A  G  G  A  C  -  U  C  U  U  C
A  A  U  C  A  C  C  C  G  A  U  -
```

- Step 2: **correlated positions** detection

```
G  A  G  C  -  C  C  A  G  U  U  C
-  A  G  G  A  C  -  U  C  U  U  C
A  A  U  C  A  C  C  C  G  A  U  -
```

# tRNA structure



# tRNA structure

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

# Implementation

# Approach

- **Pairwise alignment**

2 sequences → detect a syntactic similarity → Is there a common function?

- **Multiple alignment**

Family of sequences with the same function → To which syntactic conservation does this correspond?

# Score of a multiple alignment

- **Sum of pairs**

$$SP(m_i) = \sum_{1 \leq j < k \leq n} s(m_i^j, m_i^k)$$

$m_i$ = la $i$-ème colonne de l'alignement

$m_i^j$ = $j$-ème aa dans la colonne $i$

# Score of a multiple alignment

Scoring system:

$$s(x,x)=1, \quad s(x,y)=-1, \quad s(x,-)=s(-,x)=-2, \quad s(-,-)=0$$

```
          A   A   C   G   T   A   C   G   A   T   A
          A   -   C   G   T   A   -   A   A   T   G
          G   T   C   G   T   A   -   -   T   T   A
          ---------------------------------------
(1-2)     1  -2   1   1   1   1  -2  -1   1   1  -1
(1-3)    -1  -1   1   1   1   1  -2  -1  -1   1   1
(2-3)    -1  -2   1   1   1   1   0  -2  -1   1  -1
          =   =   =   =   =   =   =   =   =   =   =
         -1  -5   3   3   3   3  -4  -4  -1   3  -1  = -1
```

# Score of a multiple alignment

Scoring system:

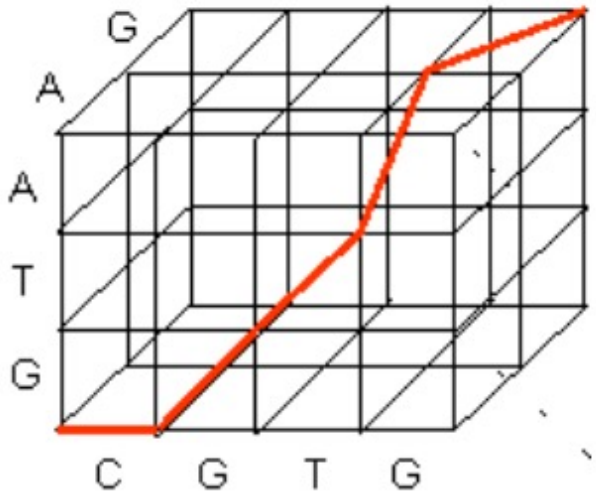
$$s(x,x)=1, \quad s(x,y)=-1, \quad s(x,-)=s(-,x)=-2, \quad s(-,-)=0$$

Alternatively… (identical)

```
      A   A   C   G   T   A   C   G   A   T   A
      A   -   C   G   T   A   -   A   A   T   G
      G   T   C   G   T   A   -   -   T   T   A
      ---------------------------------------
(1-2)  1  -2   1   1   1   1  -2  -1   1   1  -1 =  1
(1-3) -1  -1   1   1   1   1  -2  -1  -1   1   1 =  0
(2-3) -1  -2   1   1   1   1   0  -2  -1   1  -1 = -2
                                                  =
                                                 -1
```

# Exact algorithm: dynamic programming

- **Pairwise alignment** → path in a **2-dimensions matrix**
- **Multiple alignment** of *n* sequences → path in a ***n*-dimensions matrix**




```
C   G   T   -   G
-   G   T   A   -
-   -   -   A   G
```

But, impossible to use in practice… Will be **too time consuming**…

# Use of heuristics

- **Definition**: algorithm using simple rules to reduce the search space for solutions (but not necessarily giving the best solution)

- Examples: Clustal, Dialign, Muscle, Multalin, T-coffee…
  → as many programs that can produce different alignments !

- The algorithms will be seen next semester…

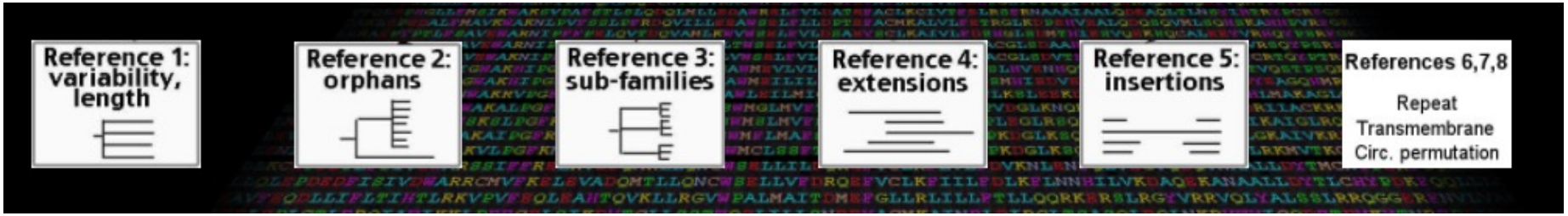
| Méthode | Idée | Stratégie |
|---|---|---|
| MSA<br>DCA | Extension de l'algorithme de Needlemen et Wunsh | Simultanée |
| Clustal<br>PIMA<br>PILEUP<br>MULTALIGN<br>Dialign | Ajout successif de séquences ou groupes de séquences | Progressive |
| Saga/Coffee<br>PRRN<br>HMMT<br>MUSCLE<br>MA-FFT | Réalignment lors de l'ajout successif de séquences ou groupes de séquences | Itérative |

# What method should I use?

- It **depends on the type of sequences** to be aligned...
- **The more divergent the sequences, the less reliable the result**
- When the identity percent is **superior to 35%** → all methods are **satisfying**
- **Clustal** tends to allow **fewer gaps than Dialign2**
  - local similarity → Dialign
  - Global similarity → Clustal

# What method should I use?

- **BaliBase**: multiple alignment database for benchmark
  - More than 150 protein families
  - Alignments based on secondary structure of proteins




- For References 1, 2 and 3: Clustal>Dialign
- For references 4 and 5: Dialign > Clustal

# Examples of alignments

# Examples

- Example from Cédric Notredame

```
GARFIELD THE LAST FAT CAT
GARFIELD THE FAT CAT
GARFIELD THE VERY FAST CAT
THE FAT CAT
```

Alignement fourni par Clustal

```
seq1    GARFIELDTHELASTFA-TCAT
seq2    ----GARFIELDTHEFA-TCAT
seq3    GARFIELDTHEVERYFASTCAT
seq4    ------------THEFA-TCAT
```

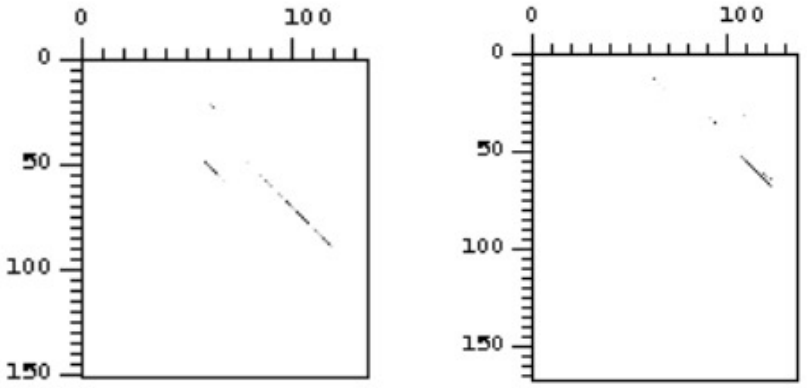
Alignement fourni par Dialign2

```
seq1    GARFIELD THE LAST FA-T CAT
seq2    GARFIELD THE ---- FA-T CAT
seq3    GARFIELD THE VERY FAST CAT
seq4    -------- THE ---- FA-T CAT
```

# Examples

- **Helix – loop – helix domain**
  - 5 sequences
  - different lengths
  - Local similarity

| | | |
|---|---|---|
| 1) | HEN1-Human | 133 |
| 2) | CBF1-Yeast | 351 |
| 3) | HES5-Mouse | 167 |
| 4) | INO4-Yeast | 151 |
| 5) | ESC1-Yeast | 413 |



# Examples

- **Clustal** result (in pink, the domain that should be aligned)

```
-------------------------------------------MMLNSDTMELD---------LPPTHSETESG------FSDCGGG
--MNSLANNNKLSTEDEEIHSARKRGYNEEQNYSEARKKQRDQGLLSQESNDGNIDSALLSEGATLKGTQSQYESG------LTSNKDE
MSSYALPSMQPTPTSSIPLRQMSQPTTSAPSNSASSTPYSPQQVPLTHNSYPLSTPSSFQHGQTRLPPINCLAEPFNRPQPWHSNSAAP
-------------------------------------------MAPSTVAVEMLSPKEKNRLRKPVVEKMRRDR------INSSIEQ
-------------------------------------------MTNDIKEIQTIQPGLSEIKEIKGELANVKKR------------

AGPD-------------------------------------------------------------------GAGPGG------------
KGSDDEDASVAEAAVAATVNYTDLIQGQE---------DSSDAHTSNQTNANGEHKDSLNGERAITPSNEGVKPNTSLEGMTSSPMEST
ASSSPTSATLSTAAHPVHTNAAQVAGSSSSYVYSVPPTNSTTSQASAKHSAVPHRSSQFQSTTLTPSTTDSSSTDVSSSDSVSTSASSS
LKLL---------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------

---------PGGGQARGPEPGEPGRKD------------LQHLSREERRRRRAT---------AKYRTA-------------------
QQSKNDMLIPLAEHDRGPEHQQDDEDNDDADID-----LKKDISMQPGRRGRKPTTLATTDEWKKQRKDS-------------------
NASNTVSVTSPASSSATPLPNQPSQQQFLVSKNDAFTTFVHSVHNTPMQQSMYVPQQQTSHSSGASYQNESANPPVQSPMQYSYSQGQP
---------LEQEFARHQPNSKLEKAD------------ILEMAVSYLKHSKAFAAAAGPKSLHQDYSEG-------------------
-----------KRRSKKINKLTDGQIR------------INHVSSEKKRRELERAIFDELVAVVPDLQPQ-------------------

----------------------HATRERIRVEAFNLAFA--ELRKLLPTLPP-----------------------------DKKLSKIEILR
----------------------HKEVERRRRENINTAIN--VLSDLLPVRESSKAAILARAAEYIQKLKETDEANIEKWTLQKLLSEQNASQ
FSYPQHKNQSFSASPIDPSMSYVYRAPESFSSINANVPYGRNEYLRRVTSLVPNQPEYTGPYTRNPELRTSHKLAERKRRKEIKELFDDLKDA
----------------------YSWCLQEAVQFLTLHAASDTQMKLLYHFQRPP-------------------------APAAPAKEPPA
--------------------------ESRSELIIYLKSLSYLSWLYERNEKLR-------------------------KQIIAKHEAKT

LAIC-------------YISYLNHVLDV-----------------------------
LASANEKLQEELGNAYKEIEYMKRVLRKEGIEYEDMHTHKKQENERKSTRSDNPHEA
LPLDKSTKSSKWGLLTRAIQYIEQLKSEQVALEAYVKSLEENMQSNKEVTKGT----
PGAAPQPARSSAKAAAAAVSTSRQPACGLWRPW-----------------------
GSSSSSDPVQEQNGNIRDLVPKELIWELGDGQSGQ---------------------
```

# Examples

- **Dialign** result, first part

```
mml-------------------------------------------------------------------------------------
m---------------------------------------------------------------------------------NSLANNNKLS
MA-------------------------------------------------------------------------------------
MT-------------------------------------------------------------------------------------
mssyalpsmqptptssiplrqmsqpttsapsnsasstpyspqqvplthnsyplstpssfqhgqtrlppinclaepfnrpqpwhsnsaapaSSSPTSATLS

---------------NSDTMELD-----------------------------LPPTHSETESGFSDCGGGAGPDgagpggpgggqarg------------
TEDEEIHSARKRGYNEEQNYsearkkqrdqgllsqesndgnidsallsegatLKGTQSQYESGLTSNKDEKGSDdedasvaeaavaatvnytdliqgQED
-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------
TAAHPVHTNAAQVAGSSSSYVYS----------------------------VPPTNSTTSQAsakhsavphrssqfqsttltpstt----------DSS


---------------------------------PEPGEPGRK-----------------------------------------------------
SSDAHTSNQTNANGEHKDSLNGERAITPSNEGVKP-------------------------NTSLEGMTSSPMESTQQSKNdmliplaehdrg-------
-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------
STDVSSSDSVSTSASSSNASNTVSVTSPASSSATPLPNQPSQQqflvskndafttfvhsvhNTPMQQSMYVPQQQTSHSSGasyqnesanppvqspmqys
```

# Examples

- **Dialign** result, second part

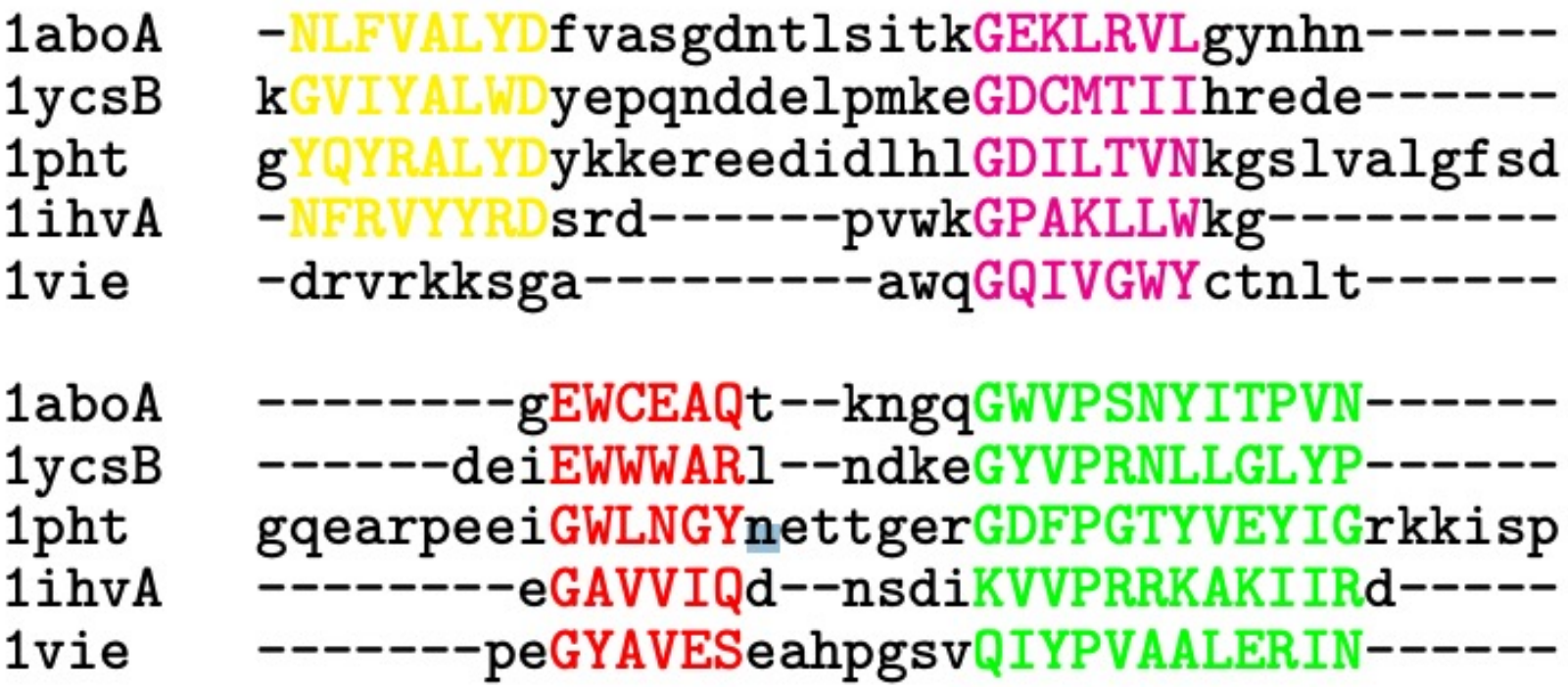
```
-----------------------------------------------------------DLQHL---SREERRRRRATA--------------K
---------PEHQqddednddadidlkkdismqpgrrgrkPTTLAttdew-KKQR-----------------------------------------
-----------------------------------------PSTVAVEMLSPKEKN-----------------------------------------
-----------------------------------------NDIKEIQTIQPGLSEIKEIKGELANVKKR---KRRSKKINKLTDG-------------Q
ysqgqpfsyPQHK-------------------------NQSFSASPIDPSMSYVYRAPESFSSINANvpyGRNEYLRRvtslvpnqpeytgpytrnpE

YRTAHATRERIRVEAFNLAFAELRKLLPTL----PPDKKLSKIEILRLAICYISYLNHVldv-------------------------------------
-KDSHKEVERRRRENINTAINVLSDLLP-V----RESSKAA---ILARAAEYIQKLKETDEanieKWTLQKLLSEQNASQLASANEKLQEELGNaykeie
-RLRKPVVEKMRRDRINSSIEQLKLLLeqefarhQPNSKLEKADILEMAVSYLKHSKAFAA----Aag-----------------------------P
IRINHVSSEKKRRELERAIFDELVAVVPDL----QPQESRSELIIYLKSLSYLSWLYERNE----KLRKQIIAKHEAKTGSSSSSDPVQEQNGNirdlvP
LRTSHKLAERKRRKEIKELFDDLKDALP-L----DKSTKSSKWGLLTRAIQYIEQLKSEQV----ALEAYVKSLEEnmqsnkevtkgt-----------

-------------------------------------------------------------------------------------
ymkrvlr--------KEGIEYEDMHThkkqenerkstrsdnphea-----------------------------------------
KSLHQDYSEGYSwclQEAVQFLTLHAasdtqmkllyhfqrppapaapakeppapgaapqparssakaaaaavstsrqpacglwrpw
KELIWELGDGQSgq-------------------------------------------------------------------------
-------------------------------------------------------------------------------------
```

# Examples

- **SH3 (Src homology 3) domain**
  - Often indicative of a protein involved in signal transduction related to cytoskeletal organization
  - 5 sequences
  - short sequences
  - low and diffuse similarity (<25%)

| | | |
|---|---|---|
| 1aboA | P00520 | 57 |
| 1ycsB | P04637 | 60 |
| 1pht | P27986 | 80 |
| 1ihvA | P00383 | 49 |
| 1vie | P12497 | 51 |

- When we **aligned secondary structures manually**



- **Clustal** result

```
1aboA   -NLFV-ALYDFVASGDNTLSITKGEKLRV-------LGYNHNG
1ycsB   KGVIY-ALWDYEPQNDDELPMKEGDCMTI-------IHREDED
1pht    -GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQ
1ihvA   ------NFRVYYRDSRD--PVWKGPAKLL----------WKGEG
1vie    ----------DRVRKKSG--AAWQGQIVGW----------YCTNL
```

```
1aboA   -------EWCEA--QTKNGQGWVPSNYITPVN------
1ycsB   EI-----EWWWA--RLNDKEGYVPRNLLGLYP------
1pht    EARPEEIGWLNGYNETTGERGDFPGTYVEYIGRKKISP
1ihvA   -------AVVIQ---DNSDIKVVPRRKAKIIRD-----
1vie    TP----EGYAVESEAHPGSVQIYPVAALERIN------
```

- **Dialign** result

Sylvain Legrand
Maître de Conférences
UMR CNRS 8198 EVO-ECO-PALEO
Evolution, Ecologie et Paléontologie
Université de Lille - Faculté des Sciences et Technologies
Bât SN2, bureau 208 - 59655 Villeneuve d'Ascq

sylvain.legrand@univ-lille.fr | http://eep.univ-lille.fr/
Tél. +33 (0)3 20 43 40 16