Recherche de similarités faibles entre séquences homologues

Limitation des comparaisons deux à deux (BLAST, FASTA, ...)

```
Seq A       CGRRLILFMLATCGECDTDSSE … HICCIKQCDVQDIIRVCC
            ::  :       :::           ::   :        :
Insuline    CGSHLVEALYLVCGERGFFYTP … EQCCTSICSLYQLENYCN
              :::   :     :  :           ::  :  :
Seq B       YQSHLLIVLLAITLECFFSDRK … KRQWISIFDLQTLRPMTA
```

Comparaisons 2 à 2:

Insuline / Seq A : 25% d'identité

Insuline / Seq B : 25% d'identité

Insulin gene family: sequence alignment

```
                B-chain                                        A-chain
INSL4   ELRGCGPRFGKHLLSYCPMPEKTFTTTPGG...[x]58 ....SGRHRFDPFCCEVICDDGTSVKLCT
INSL3   REKLCGHHFVRALVRVCGGPRWSTEA.......[x]51 ....AAATNPARYCCLSGCTQQDLLTLCPY
RLN1    VIKLCGRELVRAQIAICGMSTWS..........[x]109 ....PYVALFEKCCLIGCTKRSLAKYC
BBXA    VHTYCGRHLARTLADLCWEAGVD..........[x]25 ........GIVDECCLRPCSVDVLLSYC
BBXB    ARTYCGRHLADTLADLCF--GVE..........[x]23 ........GVVDECCFRPCTLDVLLSYCG
BBXC    SQFYCGDFLARTMSILCWPDMP...........[x]25 ........GIVDECCYRPCTTDVLKLYCDKQI
BBXD    GHIYCGRYLAYKMADLCWRAGFE..........[x]25 ........GIADECCLQPCTNDVLLSYC
LIRP    VARYCGEKLSNALKLVCRGNYNTMF........[x]58 ........GVFDECCRKSCSISELQTYCGRR
MIPI    RRGVCGSALADLVDFACSSSNQPAMV.......[x]29 ....QGTTNIVCECCMKPCTLSELRQYCP
MIPII   PRGICGSNLAGFRAFICSNQNSPSMV.......[x]44 ....QRTTNLVCECCFNYCTPDVVRKYCY
MIPIII  PRGLCGSTLANMVQWLCSTYTTSSKV.......[x]30 ....ESRPSIVCECCFNQCTVQELLAYC
MIPV    PRGICGSDLADLRAFICSRRNQPAMV.......[x]44 ....QRTTNLVCECCYNVCTVDVFYEYCY
MIPVII  PRGLCGNRLARAHANLCFLLRNTYPDIFPR...[x]86 ..EVMAEPSLVCDCCYNECSVRKLATYC
ILP     AEYLCGSTLADVLSFVCGNRGYNSQP.......[x]31 ........GLVEECCYNVCDYSQLESYCNPYS
INS     NQHLCGSHLVEALYLVCGERGFFYTPKT.....[x]35 ........GIVEQCCTSICSLYQLENYCN
IGF1    PETLCGAELVDALQFVCGDRGFYF.........[x]12 ........GIVDECCFRSCDLRRLEMYCAPLK
IGF2    SETLCGGELVDTLQFVCGDRGFYF.........[x]12 ........GIVEECCFRSCDLALLETYCATPA
            *.         .*                                    **   *    .   *
```

Descripteurs de motifs dans des séquences

- Mot exact: e.g. EcoRI site de restriction GAATTC
- Consensus: e.g. TATA box: TATAWAWR
- Expression régulière: e.g. PROSITE pattern des insulines

  C-C-{P}-x(2-4)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C
- Profils :matrice position-spécifique de pondération des substitutions et indels

Représentation d 'un motif par une *matrice de pondération position-dépendante (exemple)*

- Site donneur d 'épissage (vertébrés)
- fréquence (%) des 4 bases à chaque position
- log transformation ∏ matrice de pondération

Base	Position								
	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**33**	**60**	8	0	0	**49**	**71**	6	15
C	**37**	13	4	0	0	3	7	5	19
G	18	14	**81**	**100**	0	**45**	12	**84**	20
T	12	13	7	0	**100**	3	9	5	**46**
Cons.	M	A	G	G	T	R	A	G	T

Recherche d ’un motif dans une séquence à l ’aide d ’un profil

- Calcul des scores de similarité en faisant glisser une fenêtre de la longueur du motif le long de la séquence. Exemple:

A	33	60	8	0	0	49	71	6	15
C	37	13	4	0	0	3	7	5	19
G	18	14	81	100	0	45	12	84	20
T	12	13	7	0	100	3	9	5	46

```
GAAAGGTGAGTCAT...
GAAAGGTGA                S=18+60+8+0+0+45+9+84+15=239
.AAAGGTGAG               S=33+60+8+100+0+3+12+6+20=242
..AAGGTGAGT              S=33+60+81+100+100+45+71+84+46=620
...AGGTGAGTC             S=33+14+81+0+0+49+12+5+19=213
....GGTGAGTCA            …etc
```

Searching for distantly related homologues in sequence databases

- 1- search for homologues (e.g. BLAST)
- 2- align homologues (e.g. CLUSTAL, MEME)
- 3- compute a profile from the multiple alignment
- 4- compare the profile to a sequence database (e.g. MAST, pfsearch)

- pfsearch: http://www.isrec.isb-sib.ch/profile/profile.html
- MEME/MAST: http://meme.sdsc.edu/meme/website/

PSI-BLAST

- Position-Specific Iterated BLAST
  - 1- classical BLAST search
  - 2- compute a profile with significant BLAST hits
  - 3- BLAST search based on the profile
  - 4 -repeat steps 2-3 up to convergence

- More sensitive than Smith-Waterman
- 40 times faster

Comparison of a sequence to a database of protein motifs

- Databases: PROSITE, PFAM, PRODOM, …, INTERPRO
- Search tools:
  - ProfileScan : http://hits.isb-sib.ch/cgi-bin/PFSCAN