

M1 Bioinformatique, Connaissances et Données
Master Sciences et Numérique pour la Santé
Année 2016-2017

HMSN206 - Partie Alignement

Partie II-2 : Alignement multiple local

Anne-Muriel Arigon Chifolleau

<http://www.lirmm.fr/~arigon/enseignement/HMSN206/>



LIRMM - UM



- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- **MSA Local**
 - Méthodes à motif unique (à partir de MSA global)
 - Méthodes à alignement complet de domaines (à partir de MSA global)
 - Méthodes à plusieurs motifs (à partir de MSA global)
 - Méthodes statistiques (utilisant les méthodes de MSA local)
- Éditeurs de MSA
- Références

- Les méthodes de MSA décrites précédemment produisent un alignement global des séquences
- Bon point de départ pour les analyses phylogénétiques
- Les méthodes de **MSA locaux** alignent les **régions les plus similaires** des séquences, ignorant les régions dissimilaires
- Recherche de domaines
- Il existe de nombreuses banques de familles de séquences protéiques basées sur le partage d'une **caractéristique** commune
- La production d'un MSA local peut se faire soit
 1. en partant d'un MSA global et en sélectionnant les régions les plus conservées
 2. en utilisant des méthodes d'alignements multiple local

- Les méthodes de MSA décrites précédemment produisent un alignement global des séquences
- Bon point de départ pour les analyses phylogénétiques
- Les méthodes de **MSA locaux** alignent les **régions les plus similaires** des séquences, ignorant les régions dissimilaires
- Recherche de domaines
- Il existe de nombreuses banques de familles de séquences protéiques basées sur le partage d'une **caractéristique** commune
- La production d'un MSA local peut se faire soit
 1. **en partant d'un MSA global et en sélectionnant les régions les plus conservées**
 2. en utilisant des méthodes d'alignements multiple local

- Domaines protéiques

- Unité structurale (structuraliste) : régions capable de se replier indépendamment du reste de la protéine
- Unité fonctionnelle (biochimiste) : régions pour lesquelles une fonction propre a pu être caractérisée expérimentalement
- Unité d'évolution (évolutionniste/génomique comparative) : régions homologues, issues d'un domaine ancestral commun et conservés au cours de l'évolution

⇒ Définitions compatibles

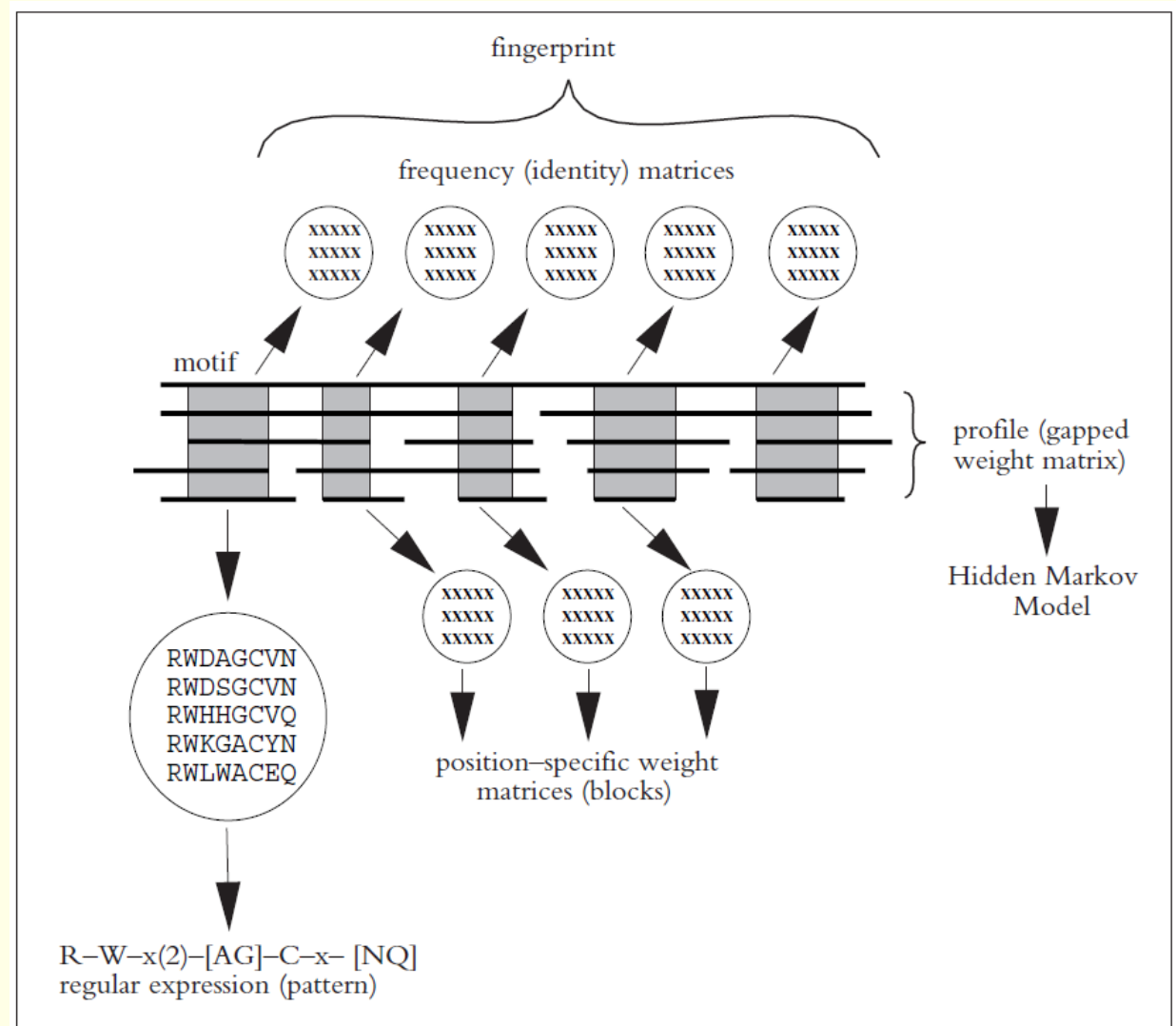
- Protéines monodomaines ou multidomaines (plusieurs fonctions ou nouvelles fonctions)

⇒ L'identification précise des domaines apporte des informations structurelles, fonctionnelles et évolutives.

- **Domaine et Motif**

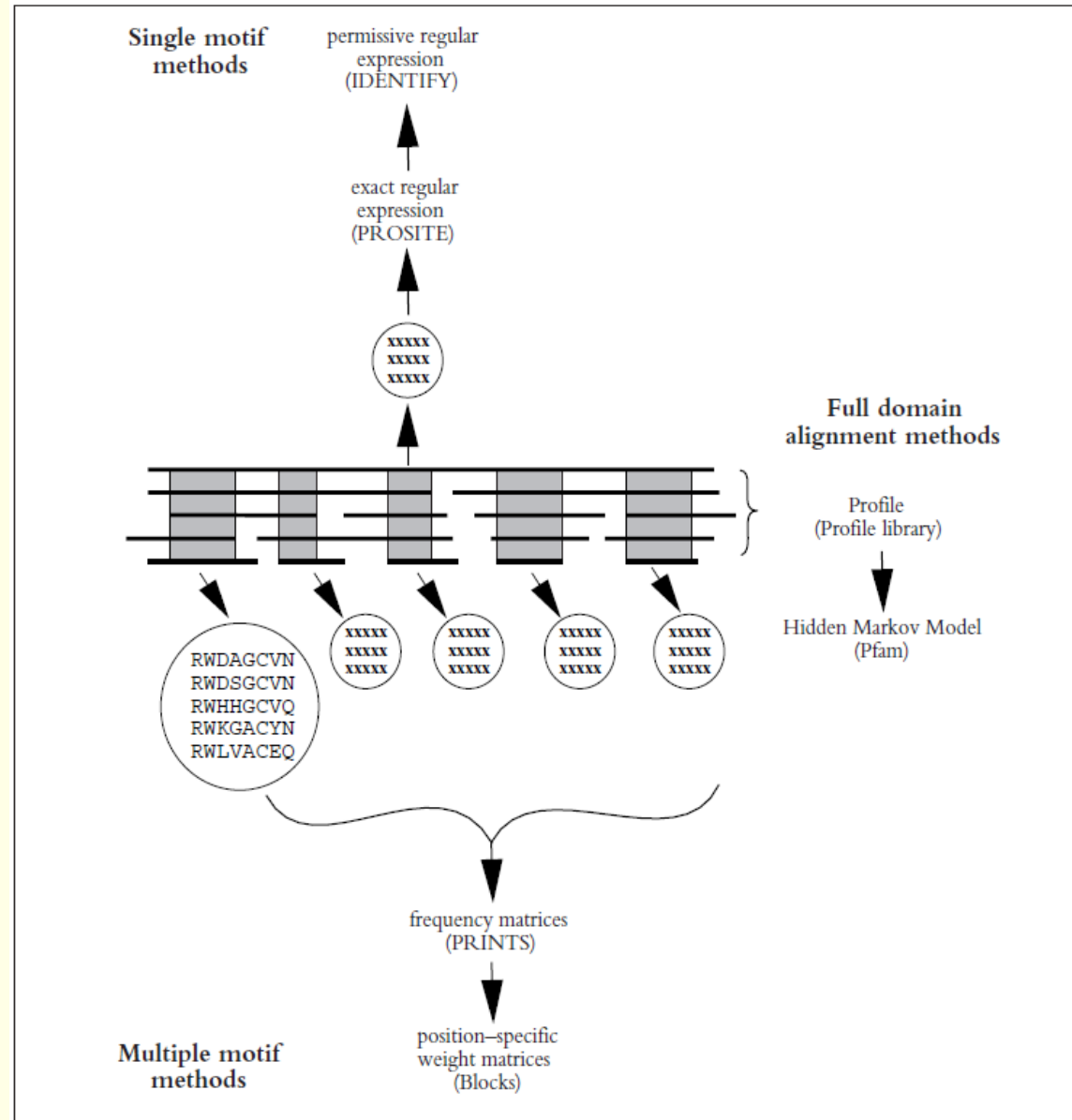
- Motif = petit groupe d'a.a. extrêmement bien conservé entre des séquences globalement différentes et associé à des interactions bien précises (site actif ou d'ancrage, ...)
⇒ Un domaine contenant plusieurs sites d'ancrage est composé de plusieurs motifs
- Un motif n'a pas forcément de repliement propre.
- Un motif contient des résidus essentiels à la fonction et à l'interaction, éventuellement entre-coupés de résidus non-essentiels

- **Motif**
 - Expression régulière / pattern
- **Groupe de motifs**
 - Matrice de fréquence
 - ⇒ Signature
 - Matrice pondérée position spécifique
 - ⇒ Bloc
- **Alignement multiple de domaines**
 - Profil
 - HMM



[Attwood, 2000]

- Méthode à motifs uniques
 - PROSITE (pattern)
 - IDENTIFY ⇒ eMOTIF
- Méthode à motifs multiples
 - PRINTS
 - BLOCKS ⇒ Interpro
- Méthode à alignement complet de domaines
 - PROSITE (profiles)
 - PFAM



[Attwood, 2000]

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
- Méthodes à motif unique (à partir de MSA global)
- Méthodes à alignement complet de domaines (à partir de MSA global)
- Méthodes à plusieurs motifs (à partir de MSA global)
- Méthodes statistiques (utilisant les méthodes de MSA local)
- Éditeurs de MSA
- Références

- La nature réutilise souvent un même élément de séquence pour réaliser des fonctions analogues
- Ces éléments conservés se structurent généralement en domaines indépendants dans les protéines et prennent en charge une des tâches de la protéine : liaison à l'ADN ou à l'ARN, fixation d'un ligand, activité enzymatique, ...
- Un gd nb de domaines sont répertoriés dans des bq spécialisées
→ Ex : PRODOM, INRA Toulouse, ~ 1 million domaines communs à au moins 2 protéines, lg moy 100 a.a.
- Alignement systématique de ces domaines $\Rightarrow$ identification de motifs caractéristiques

- Il existe des positions (d'un MSA) où l'on trouve toujours un même a.a. ou bien des a.a. similaires (F, Y aromatiques), elles constituent une **signature caractéristique** de la famille de domaines homologues
- On peut représenter un motif par un **consensus** : résidus les plus fréquents à chaque position (utilisation du code IUPAC)
- On peut représenter graphiquement un motif par une "**sequence Logo**"
- On peut avoir une représentation sous forme d'**expressions régulières** (*regex* en anglais) : syntaxe PROSITE

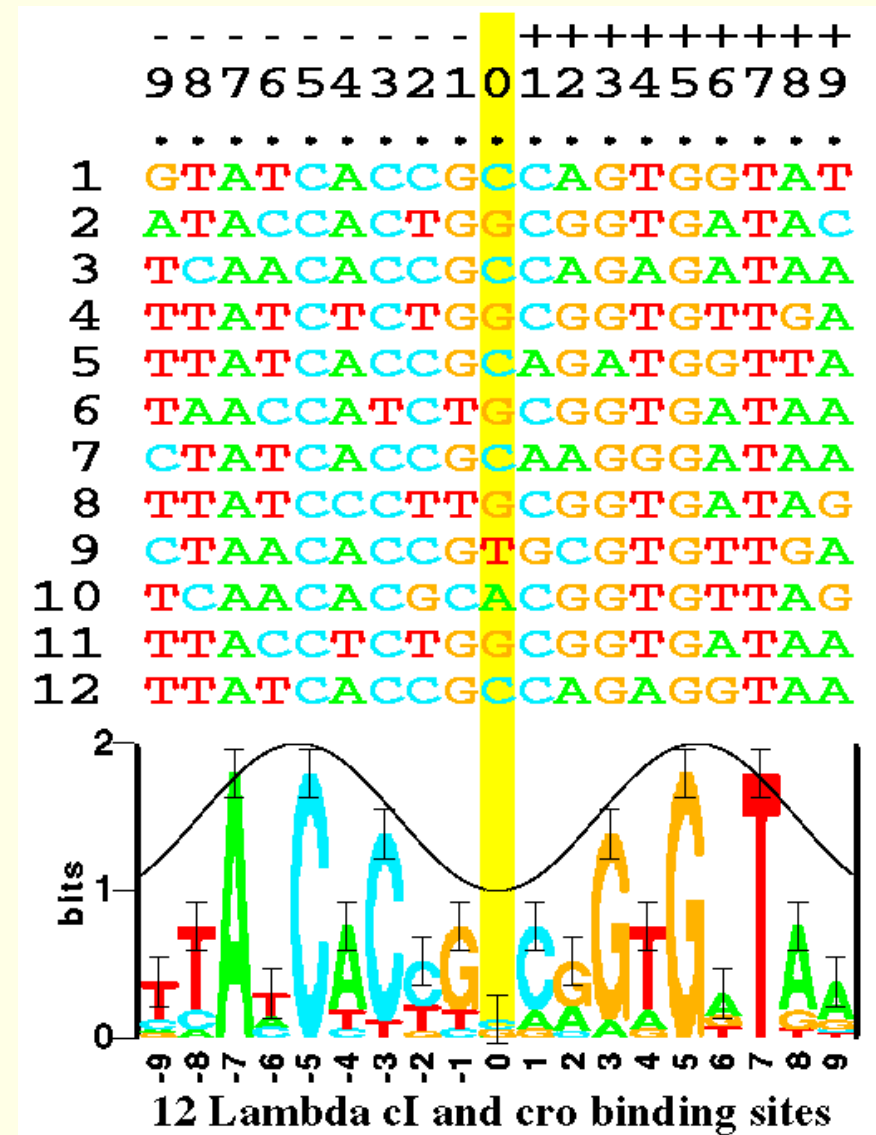
- Pour chaque position / colonne de l'alignement, on prend le résidu le plus fréquent
- Utilisation du **code IUPAC** (International Union of Pure and Applied Chemistry)

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A

Code	Description
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

- Séquence Logo [Schneider and Stephens, 1990] (WebLogo [Crooks et al., 2004])
- Représentation graphique d'un MSA
- Une pile de symboles pour chaque position
- Hauteur de la pile = conservation à cette position
- Hauteur des symboles = fréquence relative de chaque base à cette position
- Description plus riche et plus précise qu'une séquence consensus

⇒ Identifier plus facilement la conservation des colonnes d'un MSA



[Shaner et al., 2003]

- Utilisation du code standard des protéines (IUPAC)
- La lettre **X** est utilisée pour identifier une position qui peut accepter n'importe quel résidu
- Les positions qui peuvent accepter plrs résidus sont indiquées par des **[]** qui renferment les résidus acceptables, ex : **[GTS]**
- Les résidus polymorphes sont indiqués par des **{ }** qui indiquent quels résidus sont impossibles
- Chaque indication de contenu est séparée de la suivante par un **-**, ex : **A-T-X-T-X-X-[GTS]**
- Les répétitions de positions ayant un contenu identique sont indiquées de la manière suivante :
A-A-A→A(3) ; T-T→T(2) ; [AT](2)→A-A ou A-T ou T-A ou T-T
X(1,3) signifie que les alternatives **X, X-X ou X-X-X** sont possibles
- Exemple : " **F** ou **Y**, suivi d'un a.a. qcq, suivi de **C**, suivi de 5 a.a. qcq, suivi de **C**" ⇒ motif PROSITE = **[FY]-X-C-X(5)-C**

- **PROSITE** est une base de données où chaque entrée décrit des domaines de protéines, des familles ou des sites fonctionnels ainsi que les motifs et profils associés
- PROSITE contient plus de 1400 motifs documentés
- Consensus PROSITE pour “Zinc finger C2H2-type”

ABC3G_LAGLA/285-305	Cfs..CaekVaeflqenpHvn1..H
ABRU_DROME/546-567	Cpk..CgkiYrsahtlrthledk.H
ACE1_TRIRE/402-424	CrepGctkeFkrpcdltkHekt..H
ACE2_SCHPO/445-467	ClyngCnkrIarkynvesHiqt..H
ACE2_SCHPO/475-495	Cdl..CkagFvrhhdldkrHlri..H
ACE2_YEAST/605-627	ClypnCnkvFkrrynirsHiqt..H
ACE2_YEAST/635-657	CdfpgCtkaFvrnhdlirHkis..H
ADNP_HUMAN/514-535	Cpy..CrstFndvekmaaHrmv.H
ADNP_MOUSE/233-254	Cpy..CrstFndvekmaaHrmv.H
...	

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

- **Avantage :**

- Des algorithmes de recherche de motifs très puissants existent (domaine *pattern matching* en informatique)
- Les expressions régulières permettent un codage simple du motif
- Les méthodes qui cherchent des motifs pointent vers des courtes régions conservées qui peuvent représenter des fonctions biologiques

- **Inconvénients :**

- Manque de souplesse : « *matche* ou *ne matche pas* », pas de mesure de similarité
- N'autorisent pas les gaps
- Souvent il n'y a pas d'analyse de la significativité du motif
- Problème si l'organisme étudié ne fait pas partie des organismes utilisés pour définir le motif ⇒ plrs motif par famille de protéines
- Si la famille de protéines a plrs régions conservées, laquelle choisir ?

- *Fuzzy regex* en anglais
- Permet de relâcher les contraintes en autorisant les substitutions d'a.a. de même groupe
- Ces groupes sont définis sur la base des partages de propriétés physico-chimiques (cf. diagramme de Venn p. 0)
- Ex : motif [AS]-D-[IVL]-G-X(4)-PG-C-[DE]-R-[FY](2)-Q
→ Motif flou
dérivé [ASCGPT]-D-[IVLM]-G-X(4)-PG-C-[DE]-R-[FYW](2)-Q
- Permet d'identifier des prot. assez éloignées mais augmente le bruit
- Banque de donnée eMOTIF

Petits	A G
Petits hydroxyle	S T
Basiques	K R
Aromatiques	F Y W
Basiques	H K R
Petits hydrophobes	V L I
Moyens hydrophobes	V L I M
Acides/Amides	D E N Q
Petits/polaires	A C G S T P

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
 - Méthodes à motif unique (à partir de MSA global)
 - Méthodes à alignement complet de domaines (à partir de MSA global)
 - Méthodes à plusieurs motifs (à partir de MSA global)
 - Méthodes statistiques (utilisant les méthodes de MSA local)
- Éditeurs de MSA
- Références

- Les profils sont construits à partir d'un MSA global des séquences, en extrayant les régions les plus conservées, donnant ainsi un MSA de plus petite taille
- Une matrice de scores pour ce MSA est alors calculée, on l'appelle **profil** (*profile* en anglais)
- Un profil est composé de colonnes contenant les scores pour les *matches*, les *mismatches* et les pénalités de gap
- Une fois calculé, le profil peut être utilisé pour chercher une séquence cible

-
- Les profils sont similaires aux matrices de scores (**PSSM**)
 - Un profil est une matrice de 23 colonnes : une pour chaque a.a., plus une colonne pour un a.a. inconnu 'Z' et 2 colonnes pour les pénalités d'ouverture et d'extension de gap
 - Le profil a autant de lignes que le MSA a de colonnes
 - Une séquence consensus est indiquée verticalement à la gauche du tableau, elle représente les a.a. les plus fréquents à chaque position
 - Les scores reflètent le nb d'occurrences de chaque a.a. dans les séquences alignées

[illegible][illegible]

- 2 méthodes principales
 - Méthode **moyenne** : les scores du profil sont pondérés par la proportion de chaque a.a. dans chaque colonne (ex : **MEME**)
 - Méthode **évolutive** : basée sur le modèle d'évolution de protéine de Dayoff, elle suppose différentes vitesses d'évolution pour chaque colonne du MSA
- Évaluation de profils : le "**ROC plot**" mesure la précision d'un profil pour discriminer entre les membres ou non-membres d'une famille de protéines lors d'une recherche de similarité dans les banques
 - ROC plot = courbe qui donne le taux de vrais positifs (positifs qui sont effectivement détectés) en fonction du taux de faux positifs (négatifs qui sont détectés).
- Alternative pour représenter les aln de domaines : **HMM profile**

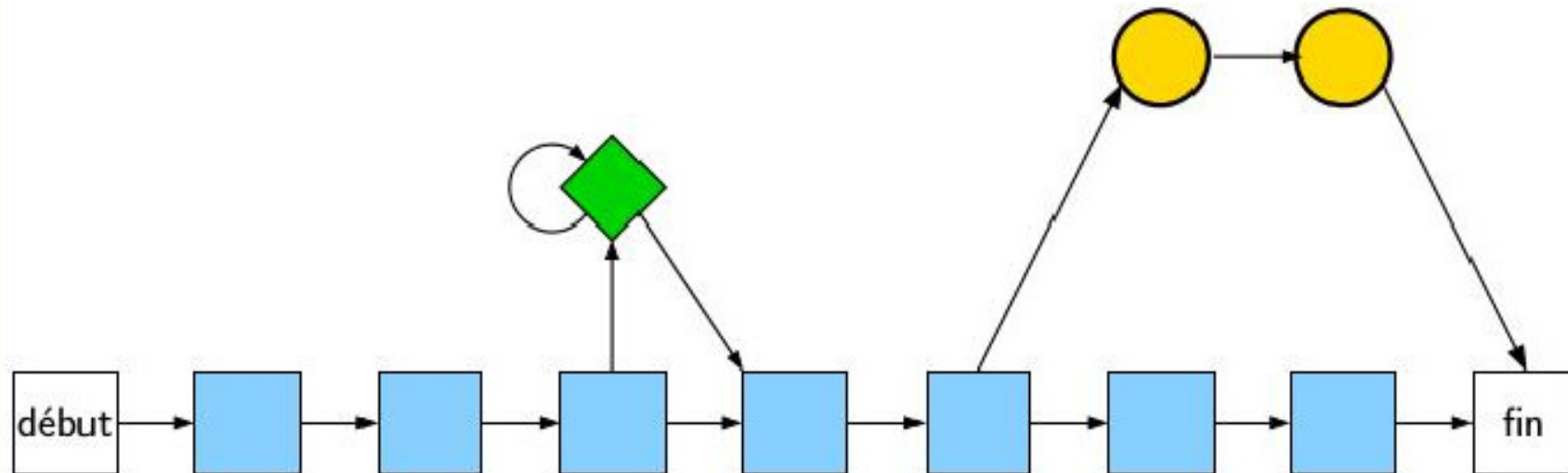
V	E	D	-	-	L	I	R	Y
V	E	D	-	-	L	R	R	Y
P	N	E	-	-	L	R	R	F
D	N	K	A	A	L	R	R	F
A	E	E	-	-	L	A	-	-



Insertion



Déletion



V E D - - L I R Y
 V E D - - L R R Y
 P N E - - L R R F
 D N K A A L R R F
 A E E - - L A - -

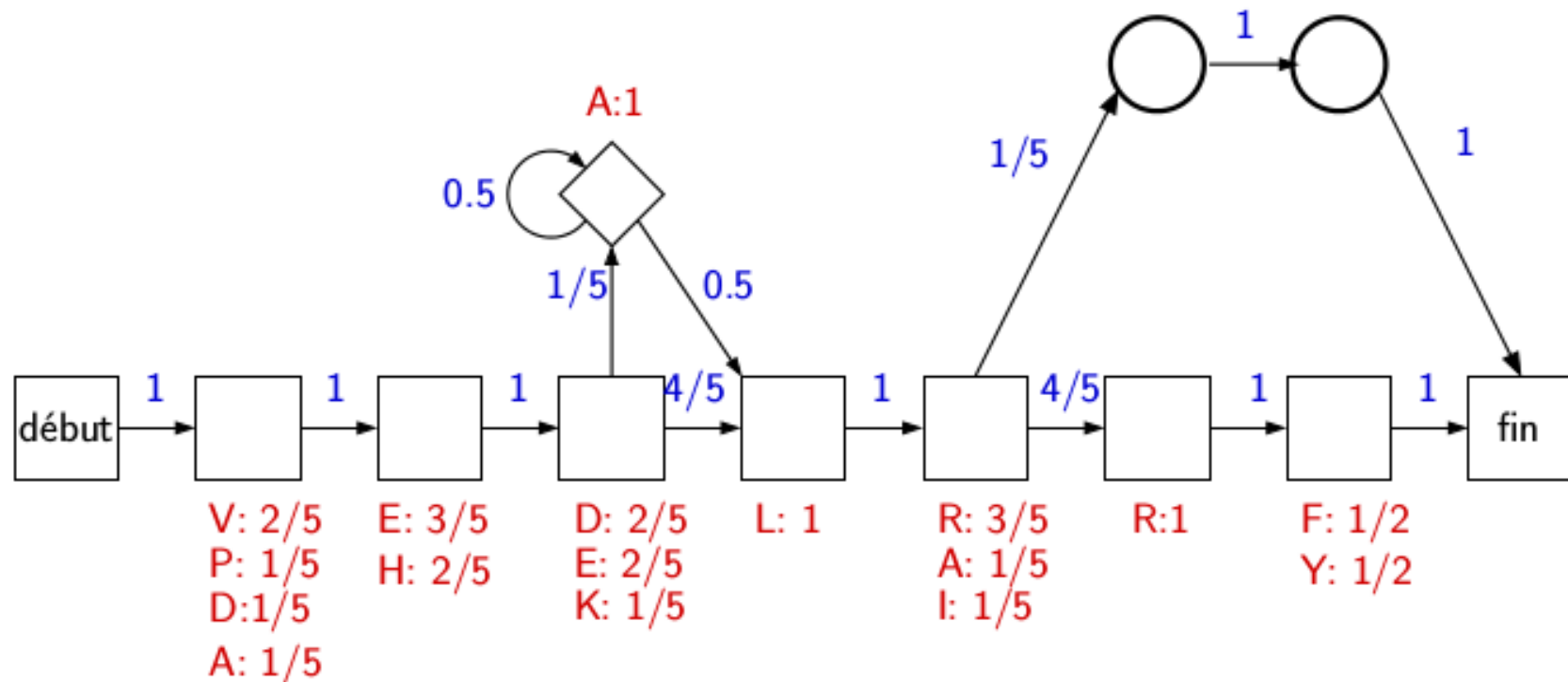
■ **émissions**

fréquences des acides aminés

■ **transitions**

circulation dans le modèle

indels



- Avantages

- Les profiles permettent de détecter des **ressemblances très faibles**
- Les profiles ne sont habituellement **pas confinés** à de courts domaines très similaires
- Les profiles sont supposés être **plus sensibles et plus robustes** que les motifs

- Inconvénients

- Les profils sont **très liés** aux MSA globaux dont ils sont extraits (reflet de la variation dans les séquences)
 - ▷ Si il y a beaucoup de séquences similaires, le MSA et le profil déduit seront **biaisés** en faveur de ces séquences
 - ▷ Diverses corrections dont pondération des séquences
- **Autre problème** : des a.a. peuvent ne pas être représentés dans certaines colonnes à cause d'un trop petit nb de seq. incluses

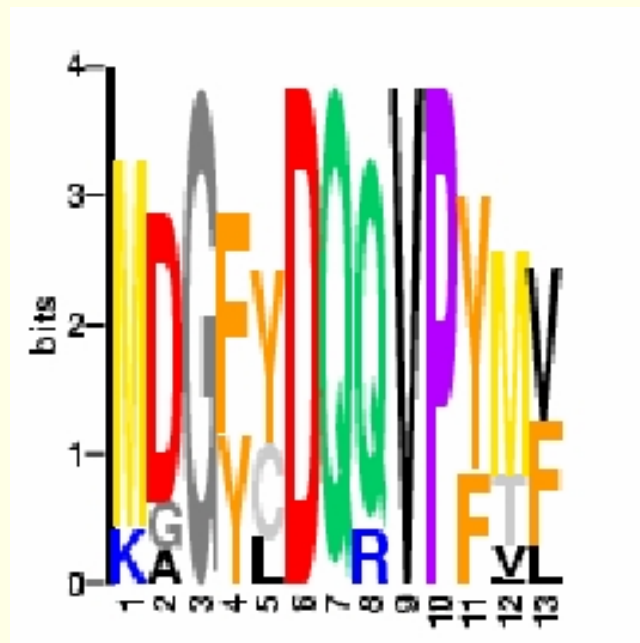
- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
 - Méthodes à motif unique (à partir de MSA global)
 - Méthodes à alignement complet de domaines (à partir de MSA global)
 - Méthodes à plusieurs motifs (à partir de MSA global)
 - Méthodes statistiques (utilisant les méthodes de MSA local)
- Éditeurs de MSA
- Références

- Comme les profils, les blocs représentent les régions conservées (motifs) dans un MSA global
- Les blocs produits de cette manière sont aussi bon (*ou moins bon*) que le MSA dont ils sont issus
- Les blocs sont représentés sous forme de matrice de score position-specific ($\sim$ profil)
- Les blocs sont différents des profils car ils n'autorisent pas les indels
- Les blocs peuvent être constitués de plusieurs régions similaires mais non adjacentes
- Le motif trouvé dans chaque séquence est de même lg, et qd les seq. sont alignées, les caractères sont dans les mêmes colonnes

Block IPB006715A

ETV5_HUMAN P41161	(1)	MDGFYDQQVPFMV	35
ETV1_HUMAN P50549	(1)	MDGFYDQQVPYMV	33
ETV1_MOUSE P41164	(1)	MDGFYDQQVPYVV	37
ETV4_HUMAN P43268	(73)	KAGYLDQQVPYTF	77
ETV4_MOUSE P28322	(81)	KGGYLDQRVPYTF	100
PEA3_BRARE Q9PUQ1	(5)	MDGYLDQQVPYTL	50

...



- Les empreintes ou signatures ou *fingerprints* sont une autre représentation de plusieurs régions conservés (motifs) dans le MSA
- Les empreintes sont des matrices de comptage d'a.a. non pondérées
- Ces matrices sont donc creuses : la plupart des cases contenant 0
- L'utilisation de ces matrices brutes est plutôt rare
- On construit donc des empreintes par itérations successives sur les banques Swiss-Prot et TrEMBL pour qu'elles contiennent plus de variabilité
- Banque de données PRINTS (Version 42.0) contient 2156 empreintes, encodant 12444 motifs uniques
- Plus restrictif que profil ou pattern : utiliser pour prédire l'occurrence de motifs similaires

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
 - Méthodes à motif unique (à partir de MSA global)
 - Méthodes à alignement complet de domaines (à partir de MSA global)
 - Méthodes à plusieurs motifs (à partir de MSA global)
 - Méthodes statistiques (utilisant les méthodes de MSA local)
- Éditeurs de MSA
- Références

- Les méthodes que l'on vient de voir utilisent un MSA global comme point de départ et tentent de localiser les régions conservées
- Une **alternative** est d'utiliser des **méthodes statistiques** pour chercher les motifs conservés sans préalablement les aligner
- Exemple de méthode :
 1. Prendre un segment de longueur L sur chacune des séquences.
 2. Choisir l'une des séquences, déplacer la fenêtre de longueur L et, pour chaque position, calculer le nouveau score de l'alignement avec les segments de longueur L des autres séquences. Choisir la meilleure position (variante : en choisir une au hasard, avec une probabilité proportionnelle à son score).
 3. Recommencer en 1.
- **Plusieurs méthodes statistiques** : algorithmes EM (*Expectation Maximisation*), Gibbs Sampler, les modèles de Markov cachés (HMM), ...
- Les matrices de scores produites par ces méthodes peuvent être utilisées pour la recherche de séquences présentant le même motif

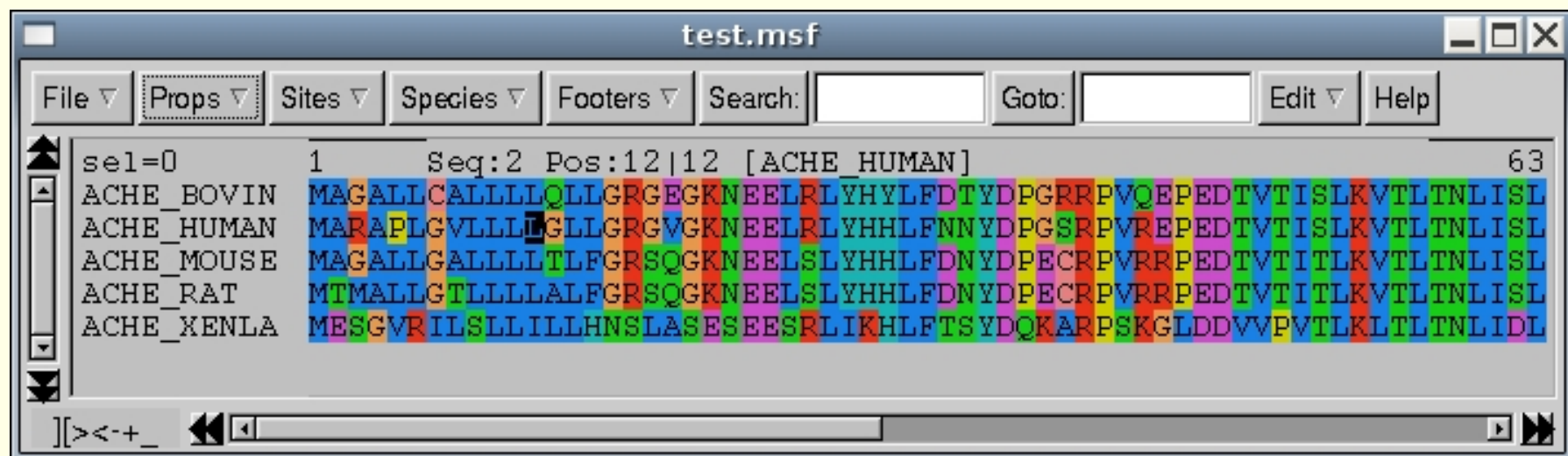
- Les bases de données comme PROSITE, PRINTS, Pfam, sont essentielles
- pour l'identification des relations entre séquences éloignées
- pour la prédiction des fonctions et des structures des protéines
- Problèmes : format différents, nomenclature différentes, avantages/inconvénients
- ⇒ InterPro ([http ://www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro))

- **InterPro** ([http ://www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro))
 - Projet mondial développée par l'EBI
 - Integrated Resource of Protein families domains and sites = vue intégrée des bd majeurs de motifs de séquences
 - Objectif = unifier les informations issues des différentes bases de données de modèles de familles protéiques.
 - BD de familles de protéines, de domaines et de sites fonctionnels
 - Unification des nomenclatures des différentes BD
 - Interface web d'accès aux données
 - Analyse d'une protéine contre Interpro : InterProScan
 - Fiche InterPro = regroupement de fiches équivalentes des différentes BD
 - Interrogation simultanée

-
- Introduction
 - Méthodes de score
 - Alignement multiple exact
 - MSA Global
 - MSA Local
 - Éditeurs de MSA
 - Références

- Une fois qu'on a obtenu un MSA global, il peut être intéressant d'*éditer* les séquences *manuellement* pour obtenir un *meilleur* alignement
- Un “*bon*” *éditeur* de séquences devrait avoir au moins les caractéristiques suivantes :
 - Prendre en compte l’affichage coloré du MSA où les couleurs *aident* à la compréhension
 - Reconnaissance des formats de MSA en sortie des logiciels d’alignement et *maintient* de ce format après l’édition
 - Interface graphique ergonomique, utilisation de la souris, ...

- La plupart d'entre eux doivent être installés sur machines



- Il existe un grand nombre de formats de MSA
- Les deux plus fréquents sont le format **MSF** (Genetic's Computer Group) et le format **ALN** (CLUSTALW)
- On peut convertir l'un en l'autre facilement
- Exemple de format ALN :

CLUSTAL W (1.82) multiple sequence alignment

```

FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA  60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA  60
*****

FOSB_MOUSE      ITTSQDLQWLQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLQPTLISSMAQSQGQPLASQPPVDPYDMPGTSYSTPGMSGYSSGGASGS 120
*****
*****.*****:*.**:*

```

- Le format **ALN** est le format d'origine du programme ClustalW. Le fichier commence par le mot clé "CLUSTAL" puis diverses informations concernant la version de CLUSTAL utilisée (ex : "CLUSTAL W (1.82) multiple sequence alignment")
- L'alignement est écrit en blocs de taille 60 résidus
- Chaque bloc commence avec le nom de la séquence et termine par la position de fin du bloc
- Des informations concernant les mises en correspondance des résidus est donnée en dessous de chaque bloc : "*" même résidu dans toute la colonne, ":" substitutions conservatives, "." substitutions semi-conservatives.

-
- Le fichier commence par autant de lignes de description que nécessaire
 - Les commentaires sont terminés par une ligne commençant avec 2 barres obliques
 - La 1re ligne reconnue est celle qui contient "MSF :", elle contient aussi la lg de la séquence, le type, etc
 - Ensuite une ligne blanche
 - Puis une ligne par séquence avec leur description (nom, lg, poids, . . .), ce bloc est aussi terminé par 2 barres obliques
 - Encore une ligne blanche
 - Enfin l'alignement proprement dit

Les commentaires

//

MSF: 510 Type: P Check: 7736 ..

Name: ACHE_BOVIN oo Len: 510 Check: 7842 Weight: 16.0

Name: ACHE_HUMAN oo Len: 510 Check: 8553 Weight: 17.8

Name: ACHE_MOUSE oo Len: 510 Check: 229 Weight: 12.5

Name: ACHE_RAT oo Len: 510 Check: 8410 Weight: 14.2

Name: ACHE_XENLA oo Len: 510 Check: 2702 Weight: 39.2

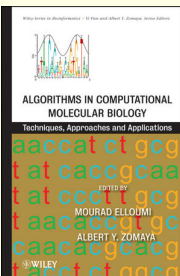
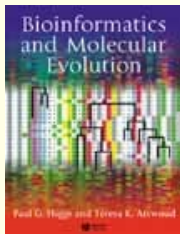
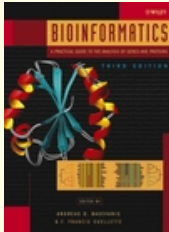
//

ACHE_BOVIN	MAGALLCALL	LLQLLGRGEG	KNEELRLYHY	LFDTYDPGRR	PVQEPEDTVT
ACHE_HUMAN	MARAPLGVLL	LLGLLGRGVG	KNEELRLYHH	LFNNYDPGSR	PVREPEDTVT
ACHE_MOUSE	MAGALLGALL	LLTLFGRSQG	KNEELSLYHH	LFDNYDPECR	PVRRPEDTVT
ACHE_RAT	MTMALLGTLL	LLALFGRSQG	KNEELSLYHH	LFDNYDPECR	PVRRPEDTVT
ACHE_XENLA	MESGVRILSL	LILLHNSLAS	ESEESRLIKH	LFTSYDQKAR	PSKGLDDVVP

ACHE_BOVIN	ISLKVTLTNL	ISLNEKEETL	TTSVWIGIDW	QDYRLNYSKG	DFGGVETLRV
ACHE_HUMAN	ISLKVTLTNL	ISLNEKEETL	TTSVWIGIDW	QDYRLNYSKD	DFGGIETLRV
ACHE_MOUSE	ITLKVTLTNL	ISLNEKEETL	TTSVWIGIDW	HDYRLNYSKD	DFAGVGILRV
ACHE_RAT	ITLKVTLTNL	ISLNEKEETL	TTSVWIGIEW	QDYRLNFSKD	DFAGVEILRV

-
- Introduction
 - Méthodes de score
 - Alignement multiple exact
 - MSA Global
 - MSA Local
 - Éditeurs de MSA
 - Références

- Chap 8 et 13 dans “Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition”
[Baxevanis & Ouellette, 2005]
- Chap 2 et 4 dans “Bioinformatique Génomique et post-génomique”
[Dardel & Képès, 2002]
- Chap 5, 7 et 9 dans “Bioinformatics and Molecular Evolution”
[Higgs & Attwood, 2005]
- Chap 12 dans “Algorithms in computational molecular biology : Techniques, Approaches and Applications”
[Elloumi & Zomaya, 2011]



- Fiches 19 à 28 dans “Bio-informatique Principes d'utilisation des outils”

[Tagu & Risler, 2010]



- “Bioinformatique Cours et cas pratique”

[Deléage & Gouy, 2013]



- Articles de recherche :

→ Notredame C (2007) Recent Evolutions of Multiple Sequence Alignment Algorithms. PLoS Comput Biol 3(8) : e123

→ Article de recherche : Thompson at al. (2011) A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods : Current Challenges and Future Perspectives. Plos One 6(3) : e18093

→ Article de recherche : Chatzou at al. (2015) Multiple sequence alignment modeling : methods and applications. Briefings in Bioinformatics 1-15