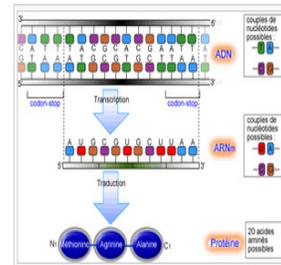


Rabat 2013 La Bioinformatique

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

@2002-2010 Sami Khuri

Bioinformatique



Sami Khuri
khuri@cs.sjsu.edu

@2002-2010 Sami Khuri

- Qu'est-ce que la Bioinformatique?
- Défis de la biologie moléculaire
- Données biologiques
- Buts de la Bioinformatique
- Bases de données
- Tâches courantes d'un biologiste
- Alignement des séquences

Qu'est-ce que la Bioinformatique? (I) Ensemble d'outils

- L'utilisation des ordinateurs pour recueillir, analyser et interpréter les données biologiques au niveau moléculaire.
- Un ensemble d'outils logiciels pour l'analyse de séquences moléculaires.



@2002-2010 Sami Khuri

Qu'est-ce que la Bioinformatique? (II)

- Discipline fondée sur les acquis de la **biologie**, des **mathématiques** et de l'**informatique**.
- Elle propose des méthodes et des logiciels qui permettent de **gérer**, d'**organiser**, de **comparer**, d'**analyser**, d'**explorer** l'information génétique et génomique stockée dans les bases de données afin de **prédire** et **produire** des connaissances nouvelles dans le domaine ainsi qu'élaborer de nouveaux concepts.

@2002-2010 Sami Khuri

Qu'est-ce que la Bioinformatique? (III)

Champs multidisciplinaire qui utilise des méthodes informatiques (mathématiques, statistiques, combinatoires...) pour :

- **Formaliser** des problèmes de biologie moléculaire
- **Développer** des outils formels
- **Analyser** les données
- **Prédire** des résultats biologiques
- **Organiser** les données.

La **bioinformatique** est une discipline relativement nouvelle, qui évolue en fonction des nouveaux problèmes posés par la **biologie moléculaire**.

@2002-2010 Sami Khuri

Bioinformatique et Données Biologiques

La **Bioinformatique** s'applique à tout type de données biologiques:

- Les **séquences** d'ADN et de protéines
- Les **structures** d'ARN et de protéines
- Les **contenus en gènes** des génomes
- Les **puces à ADN** (microarrays)
- Les **réseaux d'interactions** entre protéines
- Les **réseaux métaboliques**
- Les **arbres** de phylogénie

@2002-2010 Sami Khuri

Défis de la Biologie Moléculaire (I)

- Analyser, comprendre et organiser une grande quantité de données biologiques:
 - Des centaines de génomes complètement séquencés et enregistrés
- Projet **HapMap** du génome humain
 - Construction de la carte des haplotypes
- Projets de **séquençage** des organismes procaryotes et eucaryotes

@2009-2010 Sami Khuri

Défis de la Biologie Moléculaire (II)

- **Décoder l'information contenue dans les séquences d'ADN et de protéines**
 - Trouver les gènes
 - Différencier entre introns et exons
 - Analyser les répétitions dans l'ADN
 - Identifier les lieux des facteurs de transcription
 - Étudier l'évolution des génomes.
- **Génomique structurale:**
 - Modéliser les structures 3D des protéines et des ARN structurals
 - Déterminer la relation entre structure et fonction
- **Génomique fonctionnelle**
 - Étudier la régulation des gènes
 - Déterminer les réseaux d'interaction entre les protéines.

@2009-2010 Sami Khuri

But de la Bioinformatique

- Faire progresser les connaissances:
 - en biologie,
 - en génétique humaine,
 - en théorie de l'évolution
- Aider à la conception de médicaments
- Comprendre les maladies complexes

@2009-2010 Sami Khuri

Bases de Données Bioinformatiques

- **NCBI**, *National Center for Biotechnology Information*
 - GenBank: Séquences d'ADN
 - Site officiel de BLAST
 - PubMed: Permet la recherche de références
 - COGs: Familles de gènes orthologues ...
- **EMBL**, *The European Molecular Biology Laboratory*
- **ExPASy**, *Expert Protein Analysis System*, Protéomique
 - Swiss-Prot: Séquences de protéines
 - PROSITE: Domaines et familles de protéines
 - SWISS-MODEL: Outil de prédiction 3D de protéines
- **PDB**, *Protein Data Bank*
 - Base de données de structures 3D de protéines
 - Visualisation et manipulation de structures
- **SCOP**, *Structural Classification of Proteins*

@2009-2010 Sami Khuri

Intérêt des Séquences

- La séquence nucléotidique d'un gène détermine la séquence d'**acides aminés** de la **protéine**
- La séquence d'une protéine détermine sa **structure** et sa **fonction**
- Généralement, une similarité de séquence implique une **similarité** de **structure** et de **fonction** (l'inverse n'est pas toujours vrai).

Évolution basée, en grande partie, sur la **duplication** suivie de **modification**. D'où, beaucoup de redondance dans les bases de données.

@2009-2010 Sami Khuri

Recherche dans les Bases de Données

Tâches courantes d'un biologiste:

- Est-ce qu'une nouvelle séquence a déjà été complètement ou partiellement déposée dans les bases de données?
- Est-ce que cette séquence contient un **gène**?
- Est-ce que ce gène appartient à une **famille connue**?
Quelle est la **protéine encodée**?
- Existe-t-il d'autres **gènes homologues**?
- Existe-t-il des **séquences non codantes** similaires, des **répétitions** ou **séquences régulatrices**?

@2009-2010 Sami Khuri

Alignement de Deux Séquences

Méthode naturelle pour comparer deux séquences.
Trois opérations permises: **insertion**, **suppression**, **substitution**

Alignement Global:

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

Alignement Local:

```
CAGCAC T T - G G A T TCTCGG
      | | | | | | | |
TAGT T T A G G - T GGCAT
```

©2002-2010 Sami Khuri

Le Score d'un Alignement

Exemple: Calculons le score de l'alignement des deux séquences.

La **substitution** → -1 si les bases ne sont pas identiques,
+1 si les bases sont identiques.

La **suppression** ou **insertion** → -2

Alignement Global:

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

Score: -1 +1 -1 +1 +1 -2 +1 +1 +1 +1 +1 -2 -1 -1 +1 +1 -1 +1 -1
= 11(+1) + 6(-1) + 2(-2)
= 1

©2002-2010 Sami Khuri

BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
 - Altschul et al. 1990,1994,1997
- **BLAST** est le programme d'**alignement local** le plus utilisé.
 - Similarité locale entre une séquence donnée (requête) et une banque de données
 - Devenu populaire grâce à une implémentation très efficace.
- Les deux points forts de **BLAST** sont:
 - la vitesse
 - l'évaluation rigoureuse de la valeur statistique des alignements.

©2002-2010 Sami Khuri

La Famille BLAST

blastn: séquence donnée d'acides nucléiques comparée à des acides nucléiques d'une base de données.

blastp: séquence donnée de protéines comparée à des protéines d'une base de données.

blastx: séquence donnée d'acides nucléiques (traduite en 6 cadres de lectures) comparée à des protéines d'une base de données.

©2002-2010 Sami Khuri

Les Matrices de Substitution

- Les matrices de substitution d'acides aminés:
 - La série PAM (**P**ercent ou **P**oint **A**ccepted **M**utation)
 - Par exemple: **PAM250**
 - La série BLOSUM (**B**lock **S**UM)
 - Par exemple: **BLOSUM62** (la plus utilisée)

©2002-2010 Sami Khuri

Les Matrices de Substitution: PAM

- **PAM:** Point **A**ccepted **M**utation
 - Matrices de substitution pour les acides aminés dont les scores sont liés à la distance d'évolution
- Définition:** Deux séquences, S_1 et S_2 , divergent d'une unité PAM si la suite de mutations (substitutions) qui a converti S_1 en S_2 est telle qu'en moyenne, une seule mutation est survenue tous les 100 acides aminés.

©2002-2010 Sami Khuri

Mutations Acceptées et PAM

- **Mutations acceptées:** celles incorporées dans la protéine et transmises, sans aucun effet ou avec un effet bénéfique à l'organisme.
- Pas de correspondance absolue entre unités PAM et divergence de séquences: plusieurs mutations peuvent être survenues à la même position.
- Dans la pratique, on essaye plusieurs matrices PAM différentes. PAM 250 est la plus utilisée.

©2009-2010 Sami Khuri

Matrice de Substitution BLOSUM62

10	1	6	4	8	2	6	9	2	4	4	5	6	7	1	3	%			
A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
8	0	-4	-2	-4	0	-4	-2	-2	-2	-4	-2	-2	-2	0	0	-6	-4	A	
18	-6	-8	-4	-6	-6	-6	-2	-6	-6	-6	-6	-6	-2	-2	-4	-4	D	C	
12	4	-6	-2	-2	-6	-2	-8	-6	2	0	-4	0	-2	-6	-8	-6	D	E	
10	-6	-4	0	-6	-2	-6	-4	0	-2	4	0	0	-2	-4	-6	-4	E	F	
12	-6	-2	0	-6	0	-6	0	-6	-8	-6	-4	-4	-2	-2	-6	-4	F	G	
16	-6	-2	-4	-2	-4	0	-4	0	-2	-4	0	-2	-4	-6	-4	-4	H	H	
8	-4	2	-6	-6	-6	-4	-2	-6	-6	-2	0	-2	-4	-6	-2	-2	I	K	
10	-4	-2	0	-2	2	4	0	-2	4	0	-2	-4	-2	-4	-2	-2	L	K	
8	4	-6	-4	-4	-4	-4	-4	-2	-2	-4	-2	-4	-2	-4	-2	-2	L	L	
10	-4	-4	0	-2	-2	2	2	-2	-2	0	-4	-4	-4	-4	-4	-4	M	N	
14	-2	-4	-2	-2	-4	-8	-4	-2	-2	-4	-8	-4	-8	-4	-8	-4	N	P	
10	2	0	-2	-4	-4	-2	0	-2	-4	-4	-2	-4	-2	-4	-2	-4	O	Q	
10	-2	-2	-6	-6	-4	-4	-2	-4	-2	-4	-2	-4	-2	-4	-2	-4	R	R	
8	2	-4	-6	-4	-4	-4	-2	-4	-2	-4	-2	-4	-2	-4	-2	-4	S	S	
10	0	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	T	T	
8	-6	-2	-4	-2	-4	-4	-2	-4	-2	-4	-2	-4	-2	-4	-2	-4	V	V	
22	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	W	Y	
14																		W	Y

L'unité est le bit.

©2009-2010 Sami Khuri

PAM et BLOSUM

BLOSUM 80 BLOSUM 62 BLOSUM 45
 PAM 1 PAM 120 PAM 250

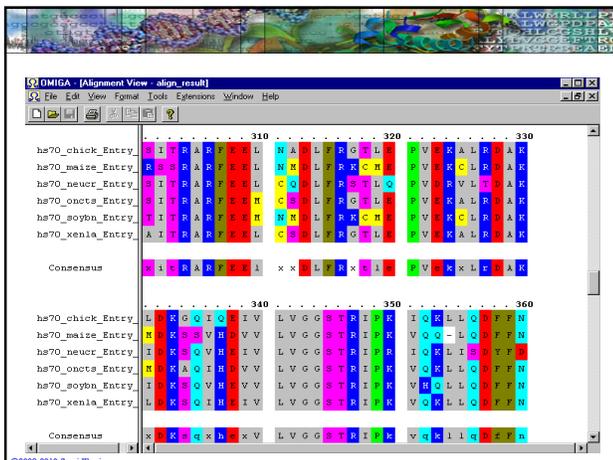
Less divergent ← → More divergent

©2009-2010 Sami Khuri

Alignement Multiple de Séquences

- Généralisation de l'alignement de 2 séquences
- Données: Un ensemble de séquences homologues (nucléotides ou acides aminés)
- Pourquoi aligner?
 - Trouver des **caractéristiques communes** à une famille de protéines
 - Relier la séquence à la **structure** et à la **fonction**
 - Caractériser les **régions conservées** et les **régions variables**
 - Dédire des **contraintes de structures** pour les ARN
 - Construire l'**arbre phylogénétique** des séquences homologues considérées.

©2009-2010 Sami Khuri



©2009-2010 Sami Khuri

Arbres de Phylogénie et Evolution

- Tous les organismes vivants dérivent d'un ancêtre commun.
- La diversité est due à la spéciation, c.à.d, à la séparation d'une espèce en deux espèces différentes.
- **Idée de base:** Les caractères sont transmis d'une génération à l'autre et, au cours de l'évolution, ces caractères subissent une série de mutations.

©2009-2010 Sami Khuri



Gènes Homologues

- **Gènes homologues:** Gènes venant d'un ancêtre commun.
- **Gènes orthologues:** Gènes homologues qui ont divergé suite à la spéciation (à la séparation d'une espèce en deux espèces différentes).
- **Gènes paralogues:** Gènes homologues qui sont issus d'un événement de duplication.

©2008-2010 Sami Khuri



Références

- Notes basées en partie sur les notes de cours de Nadia El-Mabrouk et Sylvie Hamel de l'Université de Montréal
- <http://www.infobiogen.fr/>
- <http://interstices.inria.fr/>
- www.interstices.info
- <http://www.genoscope.cns.fr/externe/HistoireBM/>
- <http://www.ac-versailles.fr/etabliss/herblay/GENETIQU/FICHES/sommaire.htm>
- <http://membres.lycos.fr/neb5000/>
- <http://www.niaid.nih.gov/factsheets/howhiv.htm>

Remerciements:
Elie Bassil, Zena Beainy, Sandra Khudeida, Nayla Klat, Myma Fayad et Yasmine Ghazzaoui ont contribué à la traduction.

©2008-2010 Sami Khuri