UNIVERSITÉ DE MONTRÉAL
Département IRO

AUTOMNE 2022
Nadia El-Mabrouk

Cours IFT3295/BIN6000/IFT6291 – Programme détaillé et sujets d'étude pour les étudiants gradués

➢ Introduction à la biologie moléculaire et à la bioinformatique (~ 3h cours, 1h démo)

- **Biologie moléculaire :** Les cellules, l'ADN, le génome, la division cellulaire, le dogme de la biologie moléculaire, l'ARN, les gènes, les protéines, transcription/traduction, séquençage de l'ADN, cartographie génétique.
- **Bio-informatique :** Définition, origine, évolution des concepts, introduction thèmes classiques (assemblage, annotation, évolution).

Sujets à explorer :

- Les algorithmes d'assemblage; Facteurs d'approximation.
  - *A. Blum, T. Jiang, M Li, J. Tromp and M. Yannakakis. "*Linear approximation of shortest superstrings*". Journal of the ACM, 41:630-647, 1994.* https://www.cs.cmu.edu/~avrim/Papers/superstring.pdf
  - *D. Breslauer, T. Jiang and Z. Jiang. "*Rotations of periodic strings and short superstrings.*" Journal of Algorithms, 24:340-353, 1997.* https://pure.mpg.de/rest/items/item_1827386/component/file_2574031/content
- Séquençage par hybridation; assemblage par plus court chemin dans un graphe (chemins hamiltoniens et eulériens).
  - *Computational Molecular Biology – An Algorithmic approach, Pavel A. Pevzner, The MIT Press, 2000, Chapitre 5.*

- Présentation simplifiée également dans : *An introduction to Bioinformatics Algorithms*, *Neil C. Jones and Pavel A. Pevzner, The MIT Press, 2004.* *https://www.cs.bgu.ac.il/~bccg131/wiki.files/An%20Introduction%20to%20Bioinformatics%20Algorithms.pdf*

➢ **Alignement de séquences (~ 6h cours, 4h Démo)**

- **Algorithmes classiques de programmation dynamique :** Algorithmes de programmation dynamique pour l'alignement global, l'alignement local, la recherche de motifs; distance d'édition, valeur de similarité; alignement local normalisé; pondération des « gaps » (trous); parallélisation de l'algorithme de programmation dynamique.
- **Mesures de similarité :** Matrices de similarité, Unités PAM, Matrices PAM, Matrices BLOSUM.
- **Optimisation des algorithmes de programmation dynamique : Algorithme de Hirschberg** pour l'alignement global en espace linéaire et application à l'alignement local; **Alignement par bande** pour une optimisation en temps et en espace de l'alignement global; Algorithme des **quatre russes** pour une amélioration asymptotique en temps de l'alignement global; Algorithme de **Crochemore et Landau** pour une amélioration asymptotique en temps de l'alignement global et local.

Sujets à explorer :

- Extension de l'algorithme de Hirschberg à l'alignement local.
  - *X. Huang, R.C. Hardison and W. Miller. "*A space-efficient algorithm for local similarities.*" Computer Applications in Biosciences, 6 (4) : 373-381, 1990.* *https://academic.oup.com/bioinformatics/article/6/4/373/356773*

- *Explorer le thème de l'alignement de génomes complets.*
  - *Evolutionary genomics, Statistical and computational methods, M. Anisimova editor, Humana Press, 2012. Chapitre 8.*

- Alignement de séquences avec réarrangements.
  - Shuffle-LAGAN pour l'alignement de paires de séquences.

*M. Brudno et al. "*Glocal alignment: finding rearrangements during alignment.*" Bioinformatics, Vol. 19, Supp. 1, i54-i62, 2003.*
http://www.cs.utoronto.ca/~brudno/i54.full.pdf

- Mauve pour l'alignement multiple.
  A.C.E. Darling, B. Mau, F.R. Blattner and N.T. Perna. "*Mauve: multiple alignment of conserved genomic sequence with rearrangements.*" *Genome Research*, 14(7):1394-403, 2004.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC442156/

- Autre méthode permettant de considérer les duplications segmentales :

  P. Holloway, K. Swenson, D. Ardell and N. El-Mabrouk.
  *"*Ancestral genome organization: an alignment approach.*" Journal of Computational Biology, 20(4): 280-295, 2013.*
  https://www.liebertpub.com/doi/full/10.1089/cmb.2012.0292

  B. Benzaid, R. Dondi and N. El-Mabrouk.
  *"*Duplication-Loss genome alignment: complexity and algorithm.*" Language and Automata Theory and Applications, (LATA) , LNCS 7810, pp 116-127, 2013.*
  http://www.iro.umontreal.ca/~mabrouk/Publications/lata.pdf

  O. Tremblay-Savard, B. Benzaid, B. F. Lang and N. El-Mabrouk.
  *"*Evolution of tRNA repertoires in Bacillus inferred with OrthoAlign.*" Molecular Biology and Evolution, Vol. 32, Num. 6, pp 1643-1656, 2015.*
  https://academic.oup.com/mbe/article/32/6/1643/1069310

- Alignements de séquences avec duplications en tandem.
  *G. Benson. "*Sequence alignment with tandem duplication.*" Journal of Computational Biology, Vol. 4, No. 3, 1997.*
  https://www.cs.umb.edu/~rvetro/vetroBioComp/compression/Sequence%20Alignment%20with%20Tandem%20Duplication.pdf

- Explorer les modèles probabilistes pour l'alignement de séquences – HMM et algorithme de Viterbi.
    - *Biological sequence analysis, Probabilistic models of proteins and nucleic acids, R. Durbin, S. Eddy, A. Krogh and G. Mitchison, Cambridge 1998. Chapitre 4.*

- Alignements de séquences circulaires
    - *R Grossi et al. "*Circular sequence comparison : algorithms and applications.*" Algorithms for Molecular Biology vol. 11, num: 12 (2016).* https://almob.biomedcentral.com/articles/10.1186/s13015-016-0076-6

➢ **Prédiction de structures secondaires d'ARN (~ 3h cours, 2h Démo)**

Thèmes : Introduction ARN; structures secondaires d'ARN, exemple de l'ARN de transfert; considérations thermodynamiques;

Application de la programmation dynamique : Algorithmes de Nussinov pour la prédiction du repliement qui maximise le nombre de paires de bases; Algorithme de Zuker pour la prédiction de la structure qui minimise l'énergie libre.

Sujets à explorer :

- ➢ Modèles probabilistes basés sur la représentation d'une structure secondaire d'ARN par une grammaire hors contexte stochastique.
    - *Biological sequence analysis, Probabilistic models of proteins and nucleic acids, R. Durbin, S. Eddy, A. Krogh and G. Mitchison, Cambridge 1998. Chapitres 9 (9.6) et* 10 (10.3).
- ➢ Algorithme de classification des systèmes CRISPR-Cas basée sur l'utilisation de la séquence et de la conservation de la structure secondaire de l'ARN des répétitions.
    - *S.J. Lange, O.S. Alkhnbashi, D. Rose, S. Will, R. Backofen, "*CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.*" Nucleic Acids Res. 41 (17) (2013) 8034–8044.*

➢ **Alignement multiple (~ 4h cours)**

Thèmes : Définition; pondération. Solution exacte, algorithme de programmation dynamique et accélération de Carillo & Lipman. Alignement consistant avec un arbre, arbre étoile et heuristique bornée; alignement phylogénétique. Heuristiques usuelles : méthodes progressives, ClustalW,

alignement d'une séquence avec une matrice consensus; méthodes itératives, méthodes par points d'ancrage.

Sujets à explorer :

- Au lieu de représenter le résultat d'un algorithme d'alignement multiple (comme ClustalW) par une matrice consensus, on peut le représenter comme un graphe orienté acyclique (alignement multiple partiellement ordonné). Un algorithme d'alignement basé sur cette représentation a été développée par les auteurs suivants :
  - *C. Lee, C. Grasso and M.F. Sharlow. "*Multiple sequence alignment using partial order graphs.*" Bioinformatics, 18:452-464, 2002.* *https://academic.oup.com/bioinformatics/article/18/3/452/236691*

  - *C. Grasso and C. Lee. "*Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems.*" Bioinformatics, 20:1546-1556, 2004.* *https://academic.oup.com/bioinformatics/article/20/10/1546/237238*

- Trouver un alignement multiple englobant un ensemble d'alignements partiels : « The maximum weight trace problem ». Problème NP-difficile. Heuristiques et algorithme ILP par J. Kececioglu.
  - *J. Kececioglu et al. "*A polyhedral approach to sequence alignment problems.*" Discrete applied mathematics, Vol. 104, Num 1–3, pp. 143-186, 2000.* *https://www.sciencedirect.com/science/article/pii/S0166218X00001943*
  - *TJ Wheeler, JD Kececioglu. "*Multiple alignment by aligning alignments.*", Bioinformatics 23 (13), i559-i568, 2007.* *https://scholar.google.com/citations?view_op=view_citation&hl=en&user=SKXmmxIAAAAJ&citation_for_view=SKXmmxIAAAAJ:Tyk-4Ss8FVUC*

- Autres algorithmes à explorer:
  - T-Coffee, DIALIGN, MAFFT, MUSCLE, ProbCons, etc.

- **Recherche exacte (~ 4h cours, 2h Démo)**

Thèmes : *Recherche d'un motif dans un texte :* Algorithmes naïf, Morris & Pratt, Knuth, Morris & Pratt, Boyer & Moore, Horspool, Sunday. *Recherche d'un ensemble de motifs dans un texte :* Algorithme de Aho & Corasick, index sur les mots.

Sujets à explorer :

- Optimisation des algorithmes KMP et Boyer-Moore pour des petits alphabets en considérant des q-mers pour effectuer de plus grands décalages.
  - *Zhu, Rui Feng; T. Takaoka (1987).* "On improving the average case of the Boyer-Moore string matching algorithm"*. Journal of Information Processing. 10 (3): 173–177.*
  - *Jong Yong Kim and John Shawe-Taylo. "*Fast string matching using n-n-gram algorithm*", SOFTWARE—PRACTICE AND EXPERIENCE, VOL. 24(1), 79–88,1994* *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.4628*
- Recherche d'un motif dans un ensemble de « textes » (variations génomiques, pangenome, génomes de différents individus de la même espèce contenant des différences) représenté par une structure particulière appelée « elastic-degenerate strings ».
  - Costas S. Iliopoulos, Ritu Kundo and Solon P. Pissis. "Efficient pattern matching in elastic-degenerate texts". Proceedings of LATA 2017. https://link.springer.com/chapter/10.1007/978-3-319-53733-7_9?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot
  - Giulia Bernardini, Nadia Pisanti, Solon P. Pissis and Giovanna Rosone. "Approximate pattern matching on elastic-degenerate text", Theoretical Computer Science, Volume 812, 2020, Pages 109-122 https://www.sciencedirect.com/science/article/pii/S0304397519305018#br0110

- **Recherche approchée (~ 3h cours)**

6.1 **Algorithmes vectoriels :** Méthode Shift-Add de Baeza-Yates-Gonnet pour la recherche avec mismatch; méthode Shift-Or pour la recherche exacte; méthode Shift-And pour la recherche approchée.

6.2 **Algorithmes de filtrage :** Méthode exacte, Algorithme de Baeza-Yates-Perlberg; Heuristiques pour la recherche dans les banques de données biologiques, FASTA, BLAST; BLAST optimisée, méthode de PatternHunter.

Sujets à explorer :

- Recherche de répétitions dans le génome en se basant sur l'alignement local ou la recherche de k-mers – Application au système CRISPR- Cas.
  - Famille des algorithmes PILER
    *E.W. Myers, R.C. Edgar, "*PILER: identification and classification of genomic repeats.*", Bioinformatics 21 (suppl 1) (2005) i152–i158.*
  - *R. C. Edgar. "*PILER-CR: Fast and accurate identification of *CRISPR repeats." BMC Bioinformatics. 8 (2007) 18.*
  - *C. Bland, T.L. Ramsey, F. Sabree, M. Lowe, K. Brown, N.C. Kyrpides, P. Hugenholtz. "*CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats." BMC Bioinformatics, 8 (2007) 209.
  - *A. Biswas, R.H.J. Staals, S.E. Morales, P.C. Fineran, C.M. Brown, "*CRISPdetect: a flexible algorithm to define CRISPR arrays." *BMC Genomics 17 (2016) i356.*
  - Identifier les répétitions à partir de fragments métagénomiques non assemblés :

    Algorithme CRASS : *C.T. Skennerton, M. Imelfort, G.W. Tyson.* "Crass: Identification and reconstruction of CRISPR from unassembled metagenomic data.*", Nucleic Acids Res. 41 (10) (2013) e105.*

    Algorithme MetaCRISPR : *J. Lei, Y. Sun, "*Assemble CRISPRs from metagenomic sequencing data*." Bioinformatics 32 (17) (2016) i520–i528.*

- **Structures de données et arbres des suffixes (~ 4h cours, 2h Démo)**

  Thèmes : Tables « Lookup », arbres des suffixes, tables des suffixes (structures de données les plus importante en bioinformatique); construction de l'arbre des suffixes – méthodes naïves et optimisations; arbre des suffixes généralisé; **Différentes utilisations des arbres des suffixes** : Recherche plus long facteurs communs à deux mots, répétitions maximales, plus longue extension commune; recherche avec « mismatchs », algorithme hybride.

Sujets à explorer :

- Algorithmes linéaires pour la construction d'arbres des suffixes : McCreight's Algorithm; Ukkonen's Algorithm; Weiner's Algorithm.
- Algorithmes linéaires pour la construction de tables des suffixes : Kärkkäinen and Sander's Algorithm; Ko and Aluru's Algorithm.
- Utilisation de l'arbre des suffixes pour identifier des structures d'ARN.
    - *Yair Horesh, Amihood Amir, Shulamit Michaeli and Ron Unger. "*A rapid method for detection of putative RNAi target genes in genomic data.*", Bioinformatics,19 Suppl 2:ii73-80, 2003.*
      https://pubmed.ncbi.nlm.nih.gov/14534175/
- Utilisation des arbres des suffixes pour l'alignement de génomes entiers. Algorithme MUMmer.
    - *A.L. Delcher, S. Kasif, R.D. Fleischmann and J. Peterson et al. "Alignment of whole genomes". Nucleic Acids Research, 27(11):2369-2376, 1999.*
    - *A.L. Delcher, A. Phillippy, J. Carlton and S.L.Salzberg. "*Fast algorithms for large scale genome alignment and comparison.*" Nucleic Acids Research, 30(11):2478-2483, 2002.*
      https://en.wikipedia.org/wiki/MUMmer
- Utilisation des arbres des suffixes pour la recherche de duplications en tandem.
    - *J.Stoye and D. Gusfield. "*Simple and flexible detection of contiguous repeats using a suffix tree.*" Theoretical Computer Science, Vol. 270, Issues 1–2, Pages 843-856, 2002.*
      https://www.sciencedirect.com/science/article/pii/S0304397501001219
    - *D. Gusfield and J. Stoye. "*Linear-time algorithms for finding and representing all tandem repeats in a string.*" Journal of Computer and System Sciences, 69(4):525-546. 2004.*
      https://www.sciencedirect.com/science/article/pii/S0022000004000364

- Explorer l'utilisation des tables des suffixes pour réduire l'espace de stockage.
    - *Handbook of Computational Molecular Biology, Srinivas Aluru ed., Chapman & Hall/CRC Computer and Information Science Series, 2005. Chapitre 7.*

- Avec la taille considérable des banques de données biologiques, la question se pose de savoir comment conserver les indexes, comme les arbres des suffixes, en mémoire externe. Il s'agit alors de minimiser le temps d'accès I/O.
  Handbook of Computational Molecular Biology, Srinivas Aluru ed., Chapman & Hall/CRC Computer and Information Science Series, 2005. Chapitre 35.
- Étudier la transformée de Burrows-Wheeler (BWT) : prétraitement utilisé en compression de données. Très utilisée en génomique, par exemple pour les problèmes d'alignement de lectures courtes issues des nouvelles technologies de séquençage d'ADN (ou d'ARN) ou pour des problèmes de comptage de mots (détection de répétitions).
  - *Langmead B, et al. "*Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.*" Genome Biol. 2009;10:R25.*
  - *H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform"*. Bioinformatics. 2009 Jul 15;25(14):1754-60.

- **Phylogénie (~ 8h cours, 4h Démo)**

Thèmes : Modèles d'évolution; Sélection naturelle **Inférence phylogénétique** : Méthodes de distance, distance ultramétrique, algorithme UPGMA; distance additive, algorithme Neighbor-Joining; Méthodes de parcimonie, parcimonie pondérée, méthode de Sankoff, algorithme de Fitch. Dénombrement et exploration des arbres phylogénétiques; inconsistance du modèle de parcimonie et introduction aux méthodes de maximum de vraisemblance. Pondération des arbres, méthodes de Bootstrap; Mesures de similarité/dissimilarité entre arbres**. Introduction aux réarrangements génomiques; Introduction à l'évolution des familles de gènes.**