Metagenomics 101

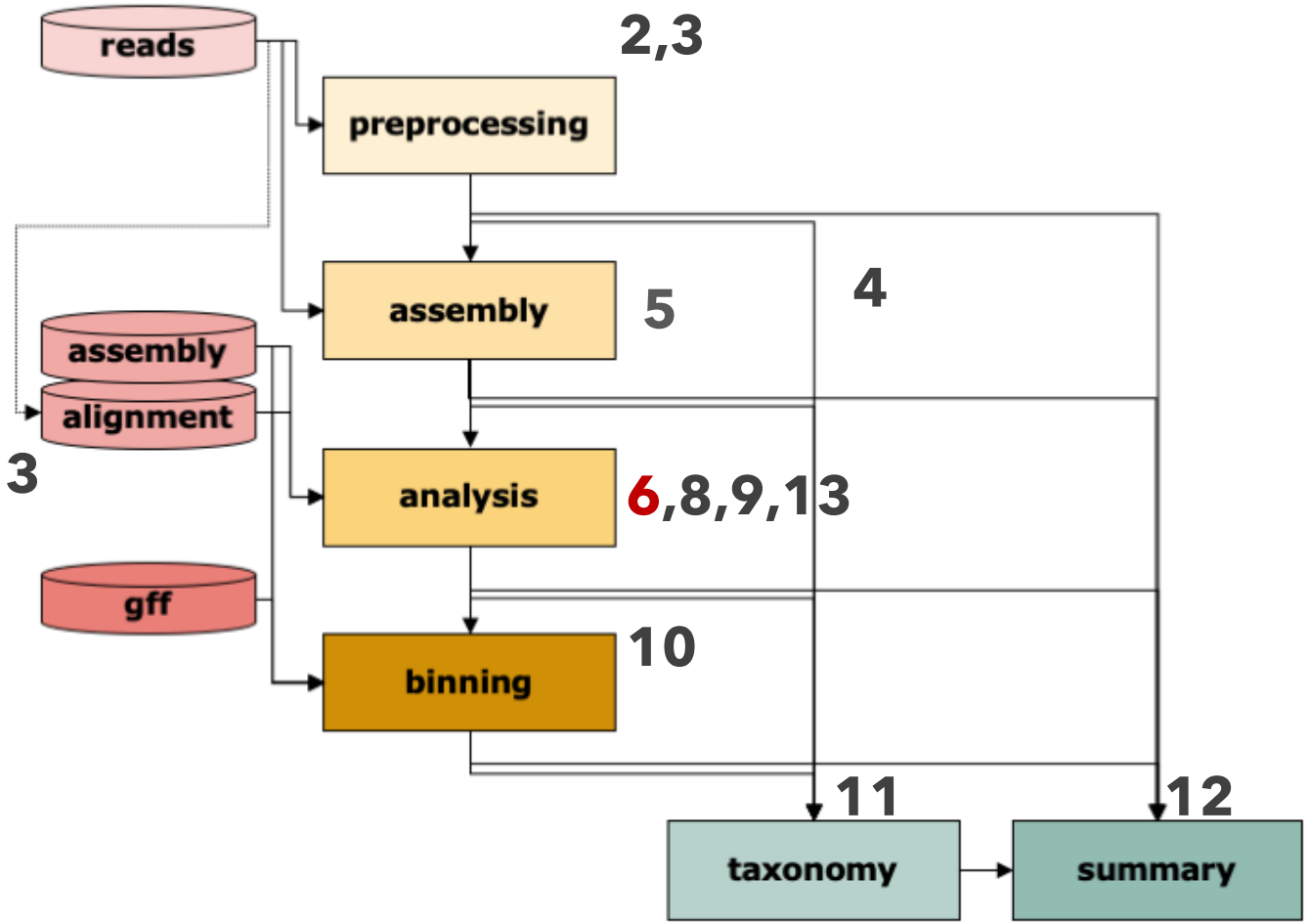
Session 6: Metagenome annotation

Anna Heintz-Buschart

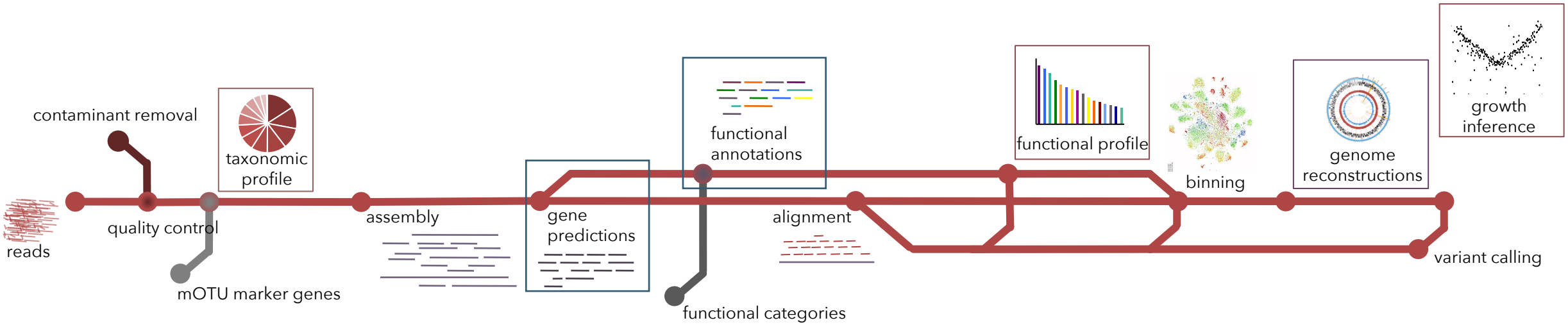
March 2022

University of Amsterdam
Life Sciences

Metagenomics (+ other omics) pipeline



Metagenomics (+ other omics) pipeline



Today

- finding bacterial genes
  - protein coding genes
  - rRNAs
- annotating genes with functions
  - why not just align?
  - HMMs and HMMER
- what we didn't look at:
  - eukaryotic genes
  - non-coding regions of interest, incl. CRISPR regions

Finding genes

- Prokka wraps tools for:
  - protein-coding genes
  - rRNA regions
  - CRISPR spacers
  - (similarity search)

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 14 2014, pages 2068–2069
doi:10.1093/bioinformatics/btu153

Genome analysis

Advance Access publication March 18, 2014

Prokka: rapid prokaryotic genome annotation

Torsten Seemann[1,2]

Seemann (2014): Prokka: rapid prokaryotic genome annotation, Bioinformatics, 30: 2068–2069

Finding protein-coding genes

Hyatt *et al. BMC Bioinformatics* 2010, **11**:119
http://www.biomedcentral.com/1471-2105/11/119

SOFTWARE Open Access

Prodigal: prokaryotic gene recognition and translation initiation site identification

Doug Hyatt[1,2*], Gwo-Liang Chen[1], Philip F LoCascio[1], Miriam L Land[1,3], Frank W Larimer[1,2], Loren J Hauser[1,3]

Hyatt et al (2010): Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119

Finding protein-coding genes

in single genomes:

- start and stop codons
- gene length
- overlaps
- bias in %GC / codon usage
- ribosomal binding sites

in a metagenomes:

- start and stop codons
- gene length
- overlaps
- bias in %GC / codon usage
- ribosomal binding sites

- pre-trained models

Hyatt et al (2010): Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119

Protein-coding genes

- positions on the contigs
- direction on the contigs
- translation

- information on completeness

.gff General feature format:

contig	source	type	start	end	score	strand	phase	attributes
contig_1001	Prodigal_v2.6.3	CDS	3	479	.	+	0	ID=GGBJBNCP_01295;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01295;partial=11
contig_1002	Prodigal_v2.6.3	CDS	3	335	.	-	0	ID=GGBJBNCP_01296;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01296;partial=11
contig_1003	Prodigal_v2.6.3	CDS	1	387	.	+	0	ID=GGBJBNCP_01297;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01297;partial=11
contig_1004	Prodigal_v2.6.3	CDS	1	1053	.	-	0	ID=GGBJBNCP_01298;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01298;partial=11
contig_1005	Prodigal_v2.6.3	CDS	2	355	.	-	0	ID=GGBJBNCP_01299;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01299;partial=11
contig_1006	Prodigal_v2.6.3	CDS	3	473	.	+	0	ID=GGBJBNCP_01300;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01300;partial=11
contig_1007	Prodigal_v2.6.3	CDS	1	849	.	-	0	ID=GGBJBNCP_01301;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01301;partial=11
contig_1008	Prodigal_v2.6.3	CDS	67	303	.	+	0	ID=GGBJBNCP_01302;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01302;partial=01
contig_1009	Prodigal_v2.6.3	CDS	1	102	.	+	0	ID=GGBJBNCP_01303;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01303;partial=10
contig_100	Prodigal_v2.6.3	CDS	2	628	.	-	0	ID=GGBJBNCP_00117;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_00117;partial=10

Annotating ‘functions’

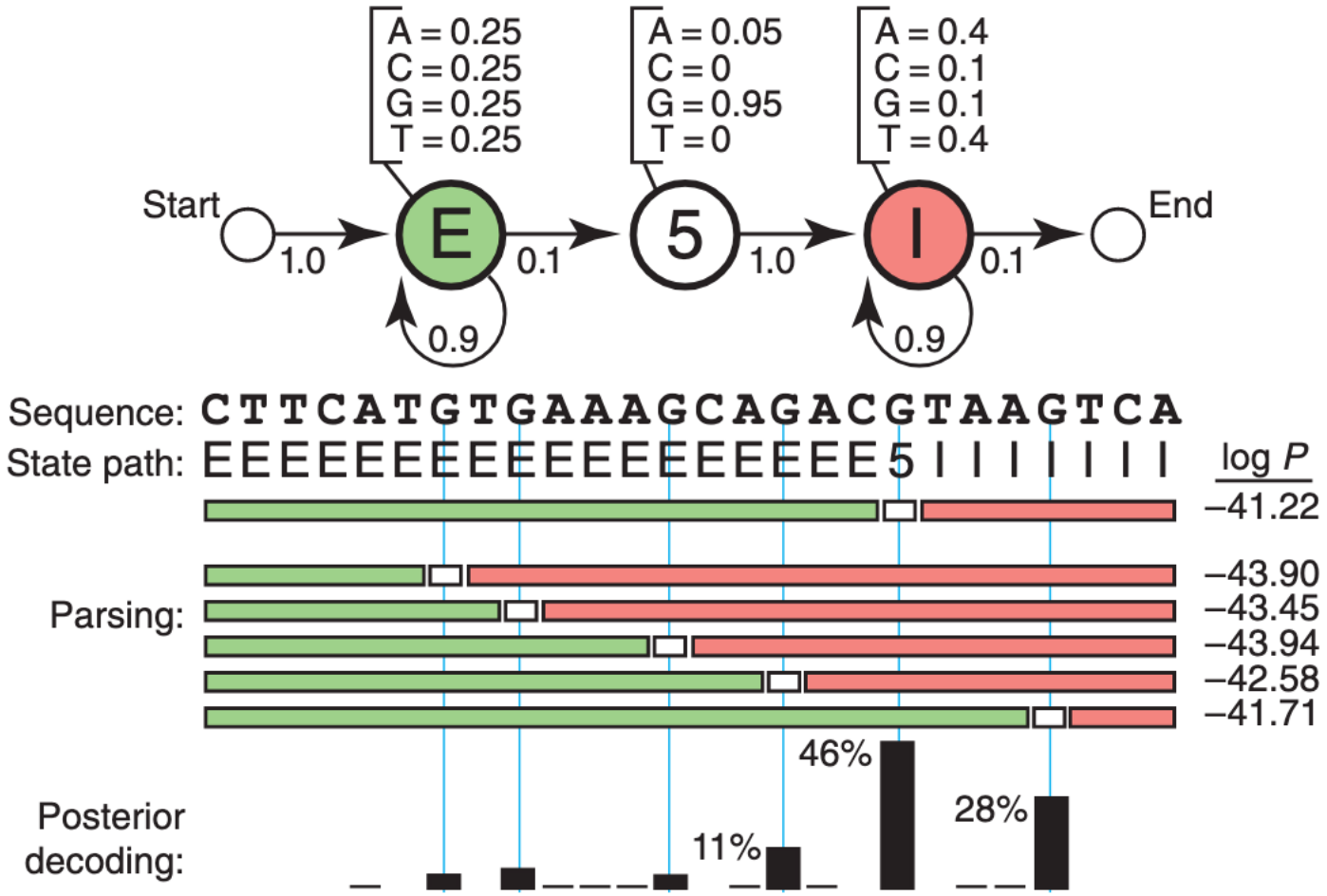
- compare the newly defined gene sequence to existing knowledge:
  - homologues in other genomes
  - classes of genes:
    - gene families
    - enzymes
    - domains
    - structures

HMMER

HMMER is a software package that provides tools for making probabilistic models of protein and DNA sequence domain families – called **profile hidden Markov models**, **profile HMMs**, or just **profiles** – and for using these profiles to annotate new sequences, to search sequence databases for additional homologs, and to make deep multiple sequence alignments. HMMER underlies several comprehensive collections of alignments and profiles of known protein and DNA sequence domain families, including the Pfam database.[1]

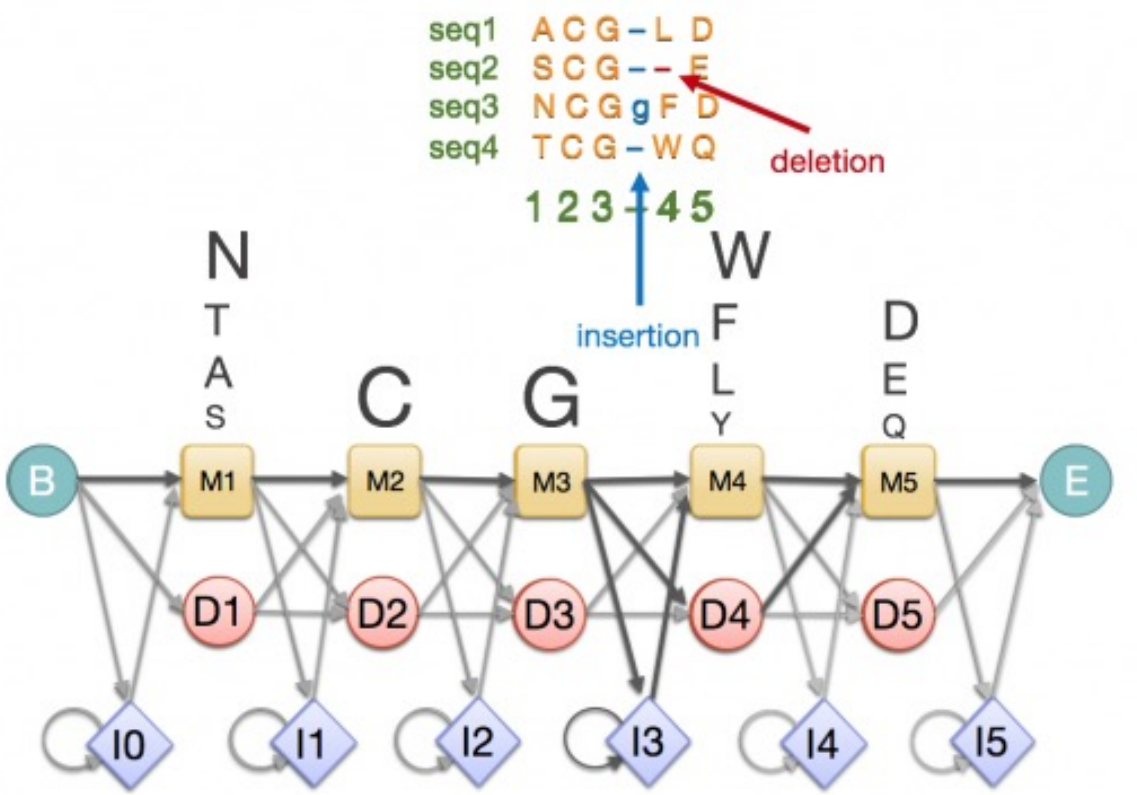
[1] pfam.org

Hidden Markov Models



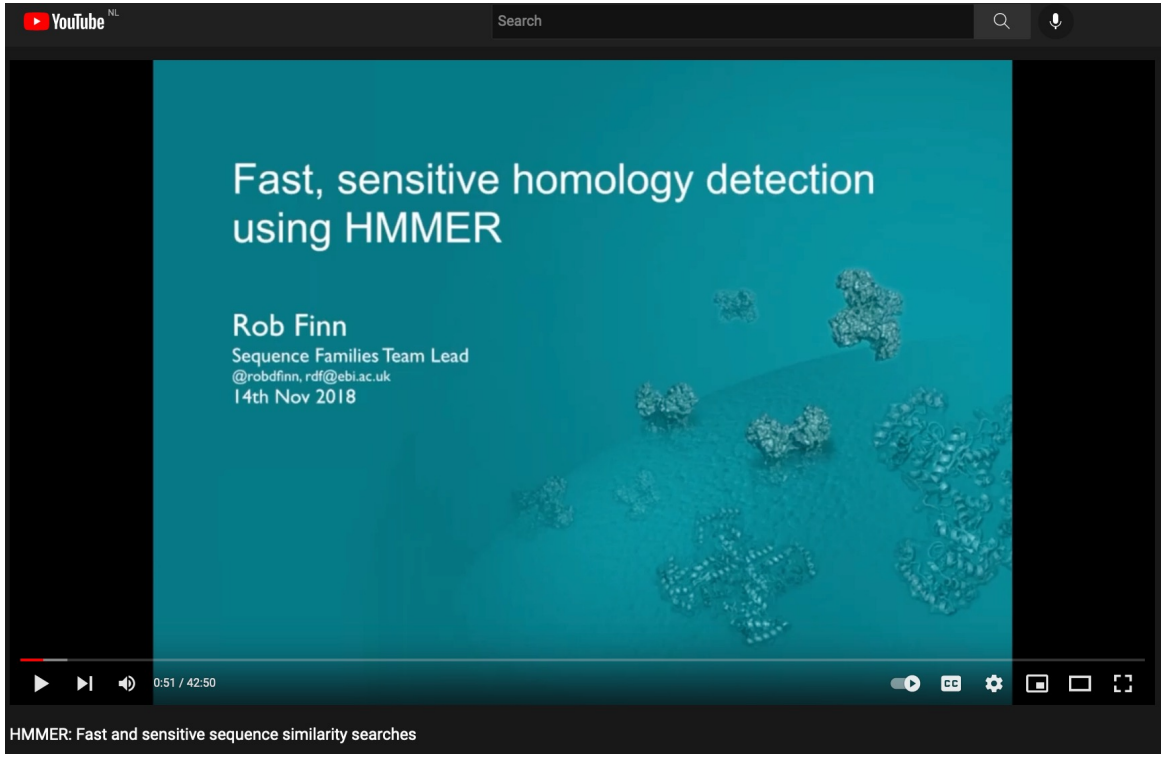
Eddy, S. What is a hidden Markov model?. Nat Biotechnol 22, 1315–1316 (2004).

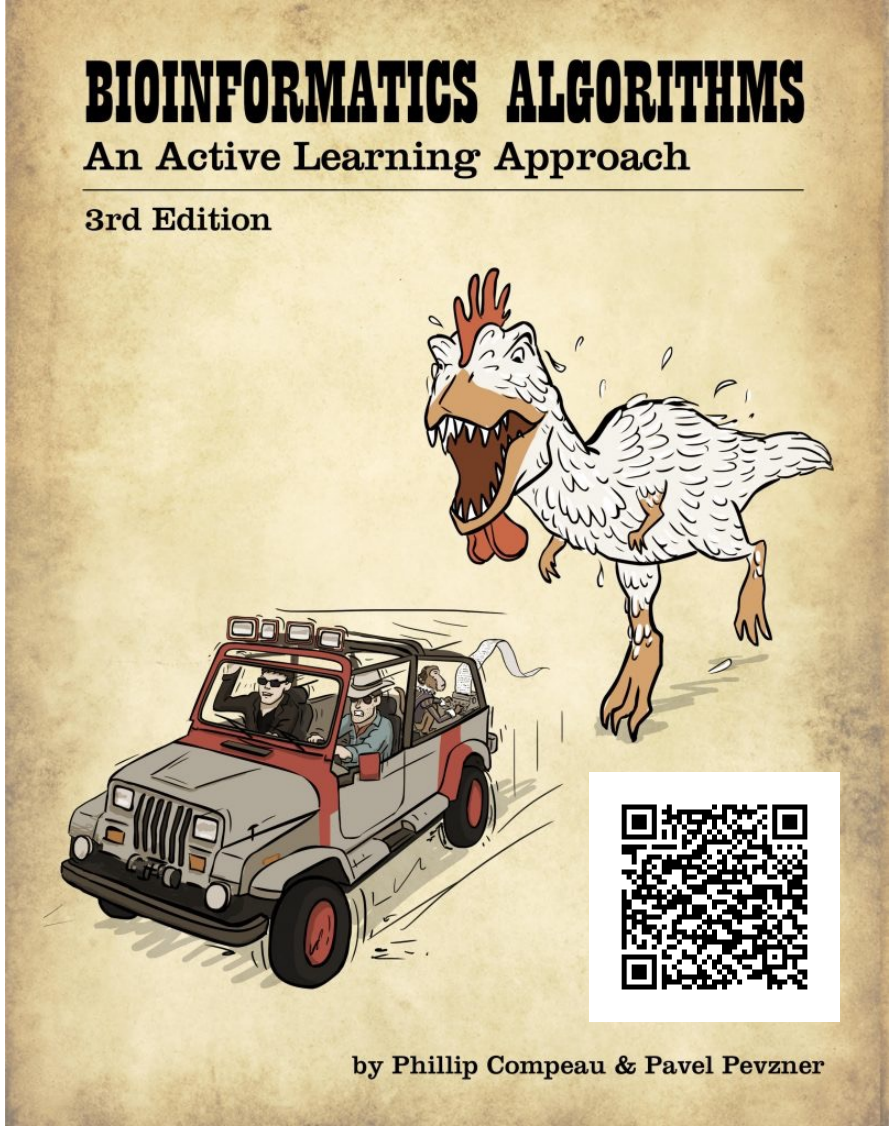
Hidden Markov Models



GCGID

more HMMs?





https://compeau.cbd.cmu.edu/online-education/bioinformatics-algorithms-an-active-learning-approach/

HMMER output

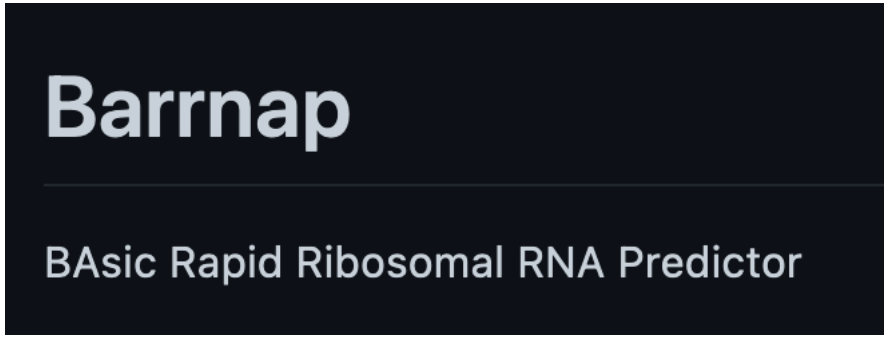
#				--- full sequence ----			--- best 1 domain ----			--- domain number estimation ----							
# target name	accession	query name	accession	E-value	score	bias	E-value	score	bias	exp	reg	clu	ov	env	dom	rep	inc	description of target
OKFJBBMB_01798	-	K00004_55	-	1.4e-57	194.1	1.5	3.8e-57	192.6	1.5	1.6	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_01232	-	K00004_55	-	5.5e-38	129.6	0.0	2.5e-37	127.5	0.0	1.8	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_01040	-	K00004_55	-	1.4e-18	65.8	0.6	1.5e-18	65.7	0.6	1.0	1	0	0	1	1	1	1	unannotated protein
OKFJBBMB_05447	-	K00004_55	-	3.6e-15	54.6	1.1	1.8e-14	52.3	1.1	1.8	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_07537	-	K00004_55	-	0.00098	17.0	0.1	0.0063	14.3	0.0	1.9	2	0	0	2	2	2	1	unannotated protein
OKFJBBMB_01132	-	K00004_55	-	0.013	13.2	0.0	0.13	10.0	0.0	2.0	2	0	0	2	2	2	0	unannotated protein
OKFJBBMB_00581	-	K00004_55	-	0.014	13.1	0.0	0.014	13.1	0.0	1.0	1	0	0	1	1	1	0	unannotated protein
OKFJBBMB_06874	-	K00007_93	-	1.1e-115	385.7	0.1	1.4e-115	385.4	0.1	1.0	1	0	0	1	1	1	1	unannotated protein
OKFJBBMB_06913	-	K00007_93	-	9e-75	250.9	0.0	9.9e-75	250.7	0.0	1.0	1	0	0	1	1	1	1	unannotated protein
OKFJBBMB_07037	-	K00007_93	-	2.3e-42	144.0	0.0	4.6e-42	143.0	0.0	1.4	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_00279	-	K00007_93	-	3.6e-32	110.4	0.0	8.4e-32	109.2	0.0	1.5	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_01323	-	K00007_93	-	2.4e-16	58.2	0.1	2.7e-16	58.1	0.1	1.0	1	0	0	1	1	1	1	unannotated protein
OKFJBBMB_00527	-	K00011_85	-	7.6e-74	247.4	0.0	5.5e-73	244.6	0.0	1.9	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_07585	-	K00011_85	-	2.7e-31	107.6	0.0	3e-31	107.4	0.0	1.0	1	0	0	1	1	1	1	unannotated protein
OKFJBBMB_01395	-	K00011_85	-	7e-31	106.2	0.0	1.5e-30	105.1	0.0	1.5	1	1	0	1	1	1	1	unannotated protein
OKFJBBMB_07910	-	K00011_85	-	8.1e-14	50.2	0.0	9.1e-14	50.0	0.0	1.0	1	0	0	1	1	1	1	unannotated protein
OKFJBBMB_04658	-	K00011_85	-	0.043	11.7	0.1	1.9	6.2	0.0	2.8	3	0	0	3	3	3	0	unannotated protein

Alternatives to HMMER

- HHblits
- alignment-based methods:
  - BLAST, DIAMOND
- many-to-many alignments:
  - MMseqs2

And what about rRNAs and tRNAs?

- rRNAs are also done using HMMs



- tRNAs:
  - first find candidates with exact match
  - then take tRNA structure into account

Nucleic Acids Research, 1994, Vol. 22, No. 11 2079–2088

RNA sequence analysis using covariance models

Sean R.Eddy* and Richard Durbin

Nucleic Acids Research, 2004, Vol. 32, No. 1 11–16
DOI: 10.1093/nar/gkh152

ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences

Dean Laslett and Bjorn Canback[1,*]

And what about …?

- what we didn't look at:
  - eukaryotic genes
  - non-coding regions of interest, incl. CRISPR regions

Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

University of Amsterdam
Life Sciences