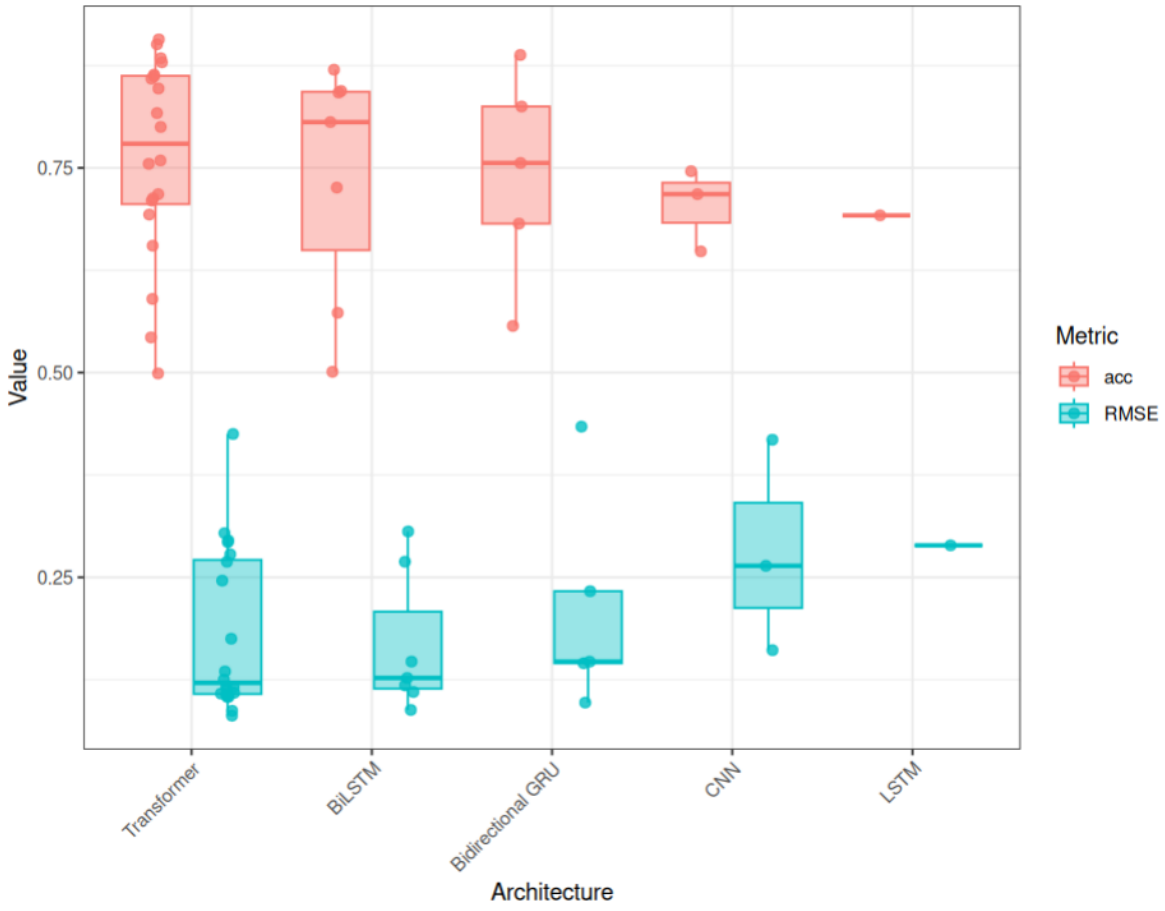
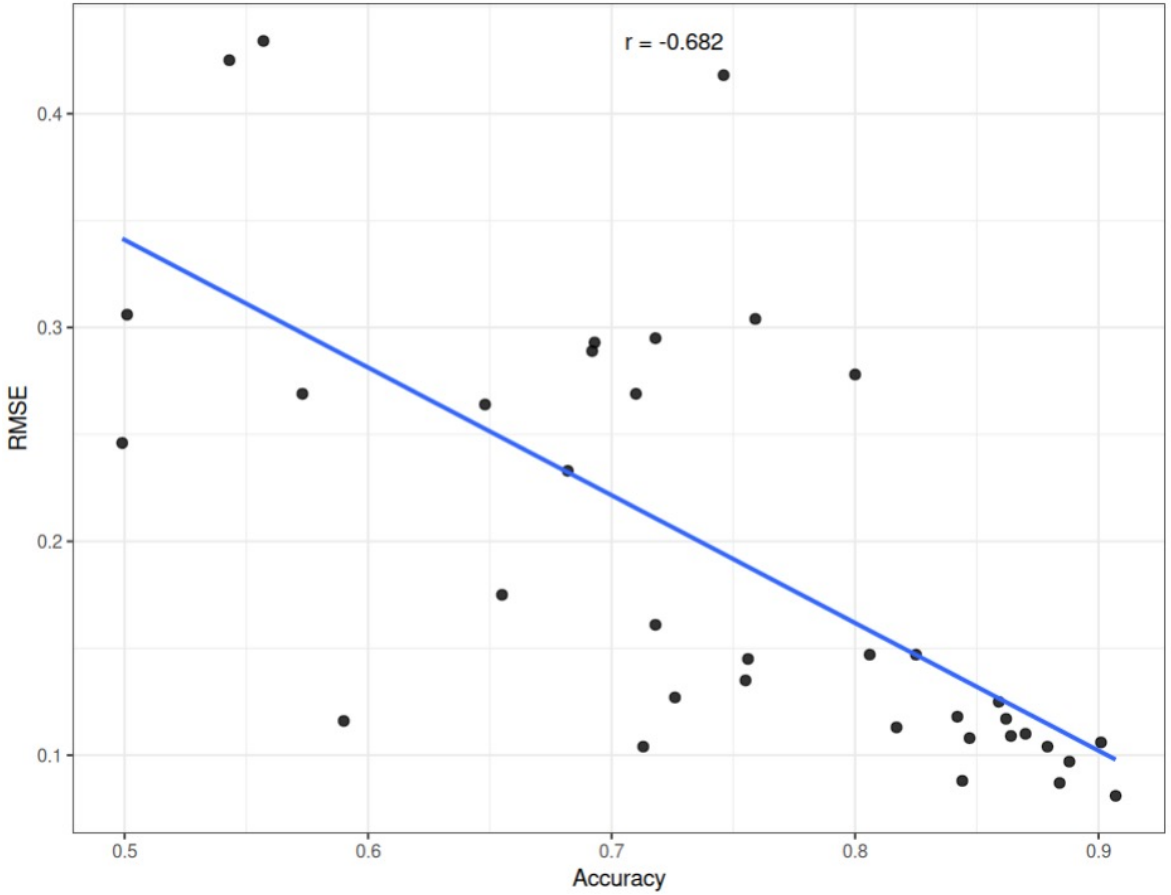
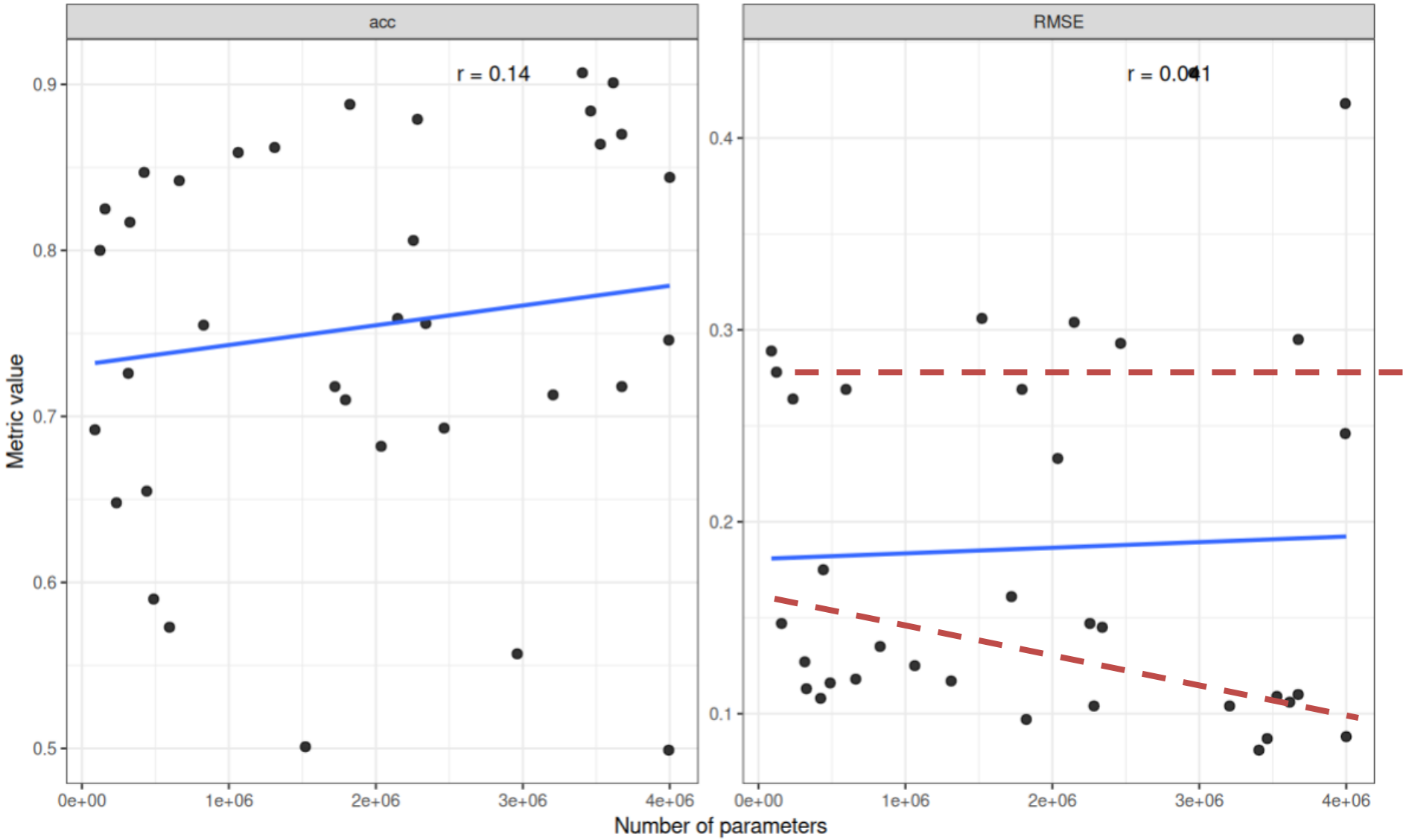
Review

Assignment 3

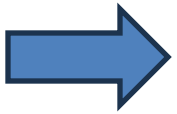




Assignment 3



Bioinformatics Problems



- Seqeunce alignment and multiple sequence alignment
- Efficient homology search (seeding method)
- Mass spectrometry based proteomics
- Gene prediction
- NGS reads mapping (efficient index structure)
- Protein structure prediction

Alignment of Protein Sequences

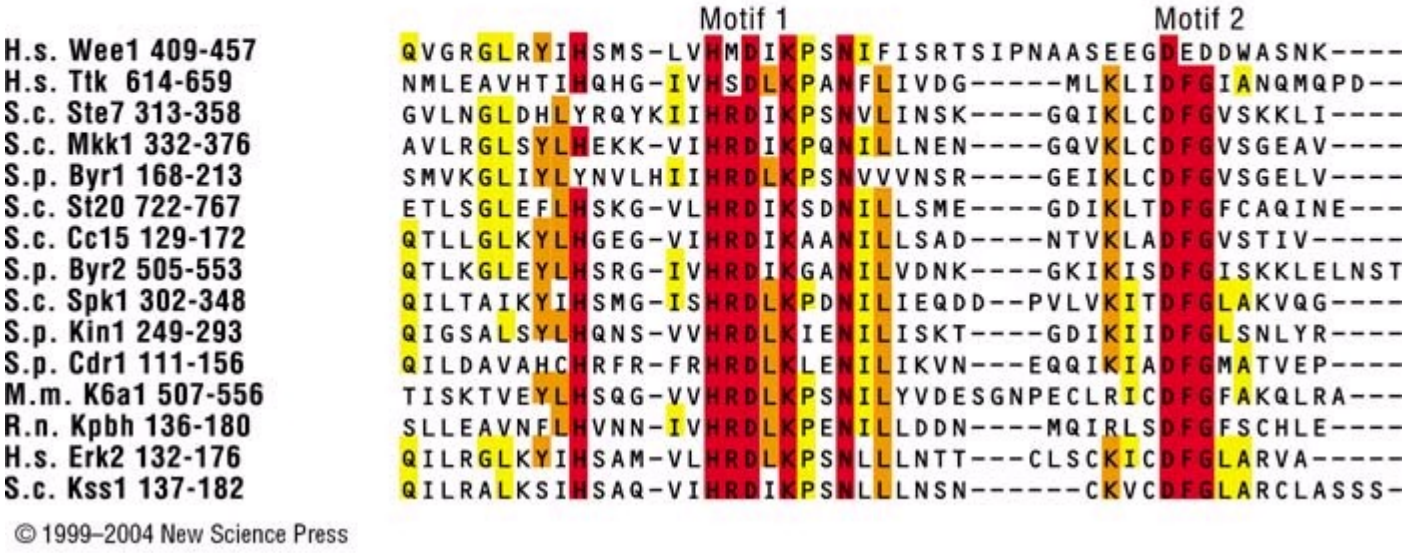
```
Conserved domain database 22426:
KOG4652, HORMA domain [Chromatin structure and dynamics]

Conserved domain length = 324 residues, 100% aligned

CT46          15  VFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLD--DLCVKILREDKNCPG--STQLVKWMLGC
                    PN +  E QSL  + RLL V++S I   RGIFPE  +  RY+D   L + +LR      G   + L K +
KOG4652       1   TLPNGLENEKQSLEFMTRLLYVAISTILRERGIFPEEYFKDRYVDGNLLVMTLLRRQDAPEGRLVSWLEKGV---

CT46          85  YDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDFISKN------QSNESSMLSTD-TKKASILL
                  +DA+++K L+ + L V T  EDP+ I E Y F F Y   G +   I+        ++ E S LS D T++    L
KOG4652       73  HDAIRQKLLKKLSL-VITESEDPEDI-EVYIFSFVYDEEGSVSARINYGINGQSSKAFELSQLSMDDTRRQFAKL

CT46         154 IRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDGDCEGVIFEGEPMYLNVGEVSTPFHIFKVKVTT
                  IRK++I  Q L PLP    +     YY E  PPDYQP GFKD           P  +N+G VSTP H   VKV
KOG4652      146 IRKLHICTQLLEPLPQ-GLILSMRLYYTERVPPDYQPEGFKDSTRAFYTLPVNPEQINIGAVSTPHHKGFVKVL-

CT46         229 ERERMENIDSTILSPKQIKTPFQKILRDKDVEDEQEHYTSDDLDIETKMEEQEKNPASSELEEPSLVCEEDEIMR
                                                 SD  D   K E
KOG4652      219 ----------------------------------------SDATDSMEKAER----------------------T

CT46         304 SKESPDLSISHSQVEQLVNKTSELDMSESKTRSGKVFQNKMANGNQPVKSSKENRKRSQHESGR---IVLHHFDS
                  K S D       V+Q +NK+ E D S S+ ++  +  N + N   PV  S+E+  +SQ   G          D
KOG4652      232 DKISDDP-FDLILVQQELNKSEEADKSFSQEKTTSITPNVLGNPLVPVDQSEEDLLKSQDSPGTGRCSCECGLDV

CT46         376 SSQESVPKRRKFSEPKEHI
                 S Q SVPK RK     EH
KOG4652      306 SKQASVPKTRKSCRKTEHG


Homology between CT46 and MGC26710 hypothetical protein

Identities = 136/249 (54%), with conservative changes = 180/249 (72%)

CT46          1   MATAQLQR-----TPMSALVFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLDDLCVKILREDK
                  MATAQL             VFP++I+ EH+SL +VK+L A S+SCITYLRG+FPE +YG R+LDDL +KILREDK
MGC26710      1   MATAQLSHCITIHKASKETVFPSQITNEHESLKMVKKLFATSISCITYLRGLFPESSYGERHLDDLSLKILREDK

CT46         71   NCPGSTQLVKWMLGCYDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDF--ISKNQSNESSMLS
                   CPGS  +++W+ GC+DAL+K+YLRM VL +YT+P   + ++E YQFKFKYT  G  MDF   S + S ES   +
MGC26710     76   KCPGSLHIIRWIQGCFDALEKRYLRMAVLTLYTDPMGSEKVTEMYQFKFKYTKEGATMDFDSHSSSTSFESGTNN

CT46         144 TDTKKASILLIRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDG-DCEGVIFEGEPMYLNVGEVST
                  D KKAS+LLIRK+YILMQ+L PLPN+V LTMKL YY+ VTP DYQP GFK+G +   ++F+ EP+ + VG VST
MGC26710     151 EDIKKASVLLIRKLYILMQDLEPLPNNVVLTMKLHYYNAVTPHDYQPLGFKEGVNSHFLLFDKEPINVQVGFVST

CT46         218 PFHIFKVKVTTERERMENIDSTIL  241
                  FH  KVKV TE  ++ ++++ +
MGC26710     226 GFHSMKVKVMTEATKVIDLENNLF  249
```

- Global alignment
- Local alignment
- Fit alignment
- Linear space alignment
- Affine gap penalty

- Loglikelihood ratio score
- BLOSUM62
- P-value
- E-value

Multiple Sequence Alignment

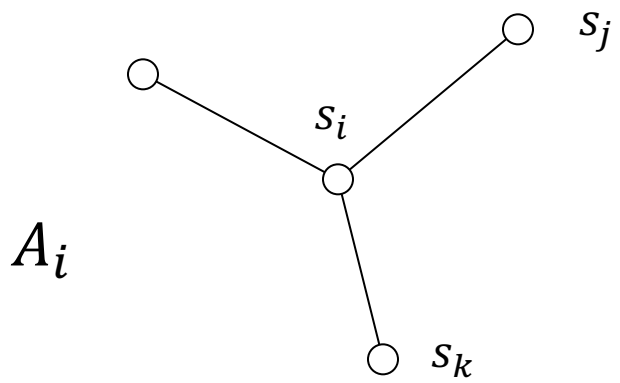
- A multiple sequence alignment of k sequences is an insertion of gaps in the positions of the sequences, just like a pairwise alignment.



- "Two homologous sequences whisper, a multiple alignment shouts loudly" -- Arthur M. Lesk

Heuristic Algorithms and Approximation Algoirthms for MSA

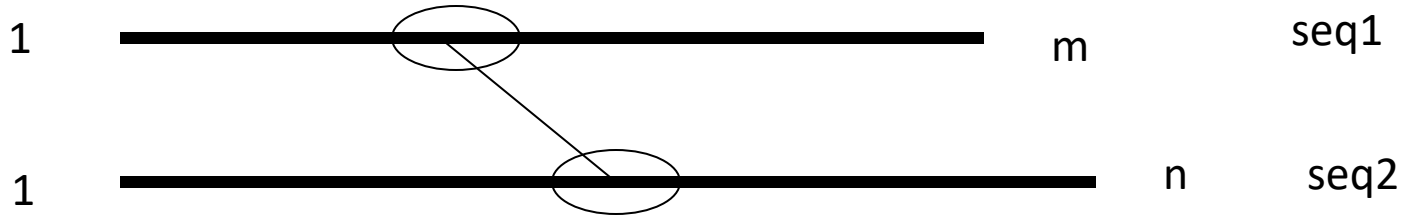
```
 t:   A-GAGC          t:   A-GA-GC
s1:   ATGAGC   ⇒     s1:   ATGA-GC   ⇒     t:   A-GA-GC
and                  and                  s1:   ATGA-GC
 t:   AGA-GC          t:   A-GA-GC         s2:   A-GTTGC
s2:   AGTTGC         s2:   A-GTTGC
```



Bioinformatics Problems

- Seqeunce alignment and multiple sequence alignment
- Efficient homology search (seeding method)
- Mass spectrometry based proteomics
- Gene prediction
- NGS reads mapping (efficient index structure)
- Protein structure prediction

Fast Homology Search



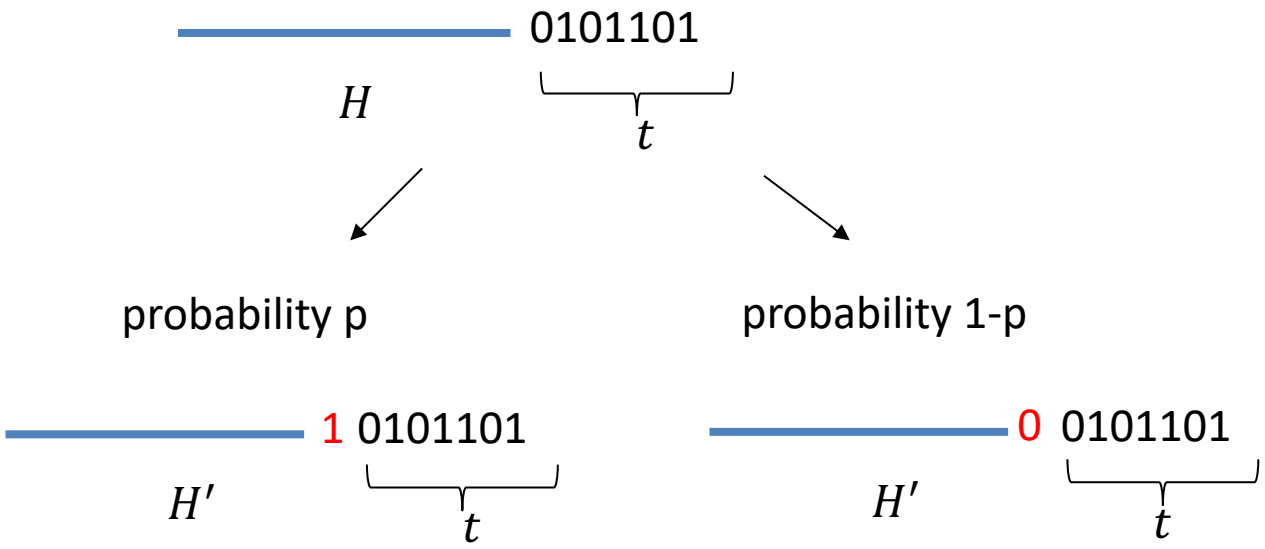
Hit and Extend

```
GAGTACTCAACACCAACATTAGTGGCAATGGAAAAT…
||  ||||||||| |||||  ||  |||||    ||||||
GAATACTCAACAGCAACACTAATGGCAGCAGAAAAT…
        111*1**1*1**11*111
```

Spaced seed

Compute Hit Probability

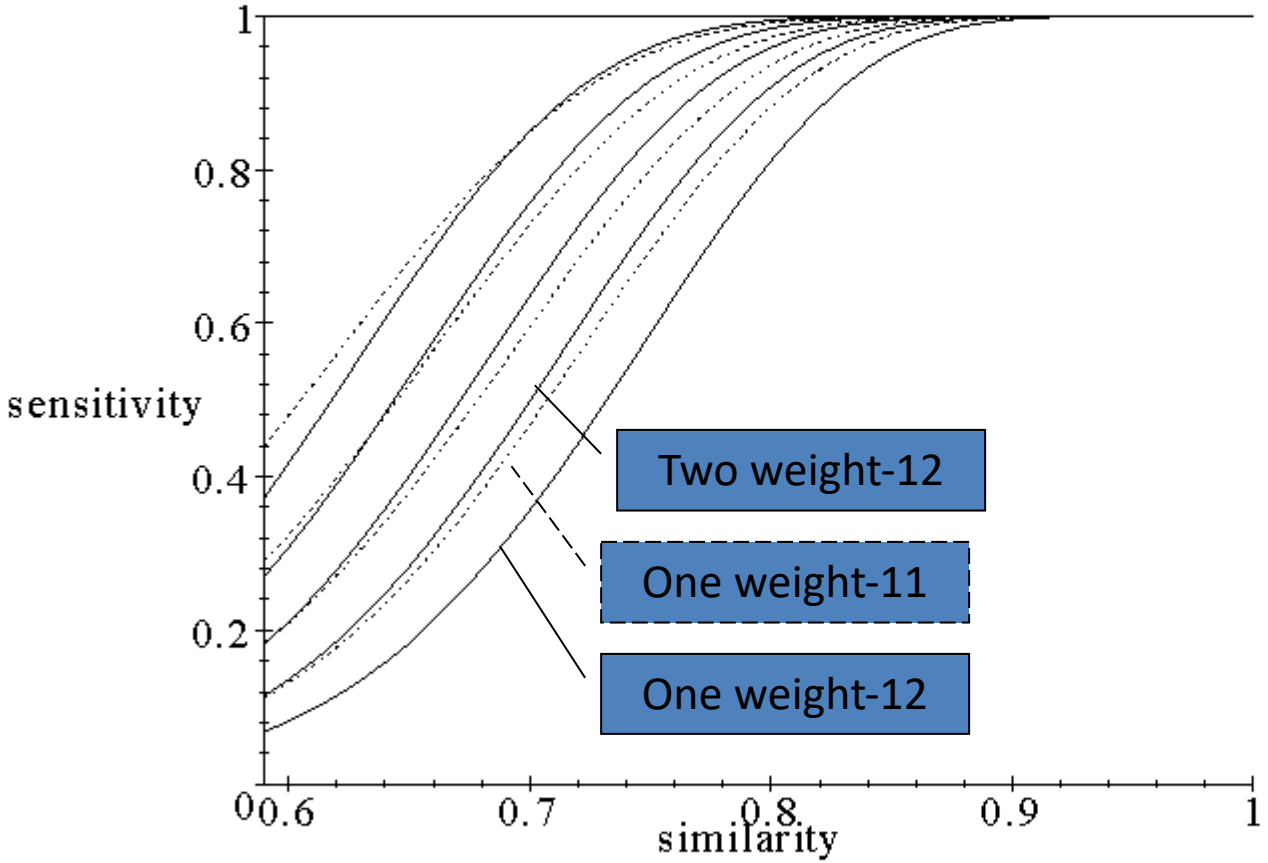
- Case I: t is hit by s. Then $D[i,t] = 1$.
- Case II: t is not hit by s:



H' is the length-(i-1) distribution. t' is the length-(l-1) prefix of t.

$$D[i,s] = p \cdot D[i-1,1t'] + (1-p) \cdot D[i-1,0t']$$

Multiple Spaced Seeds

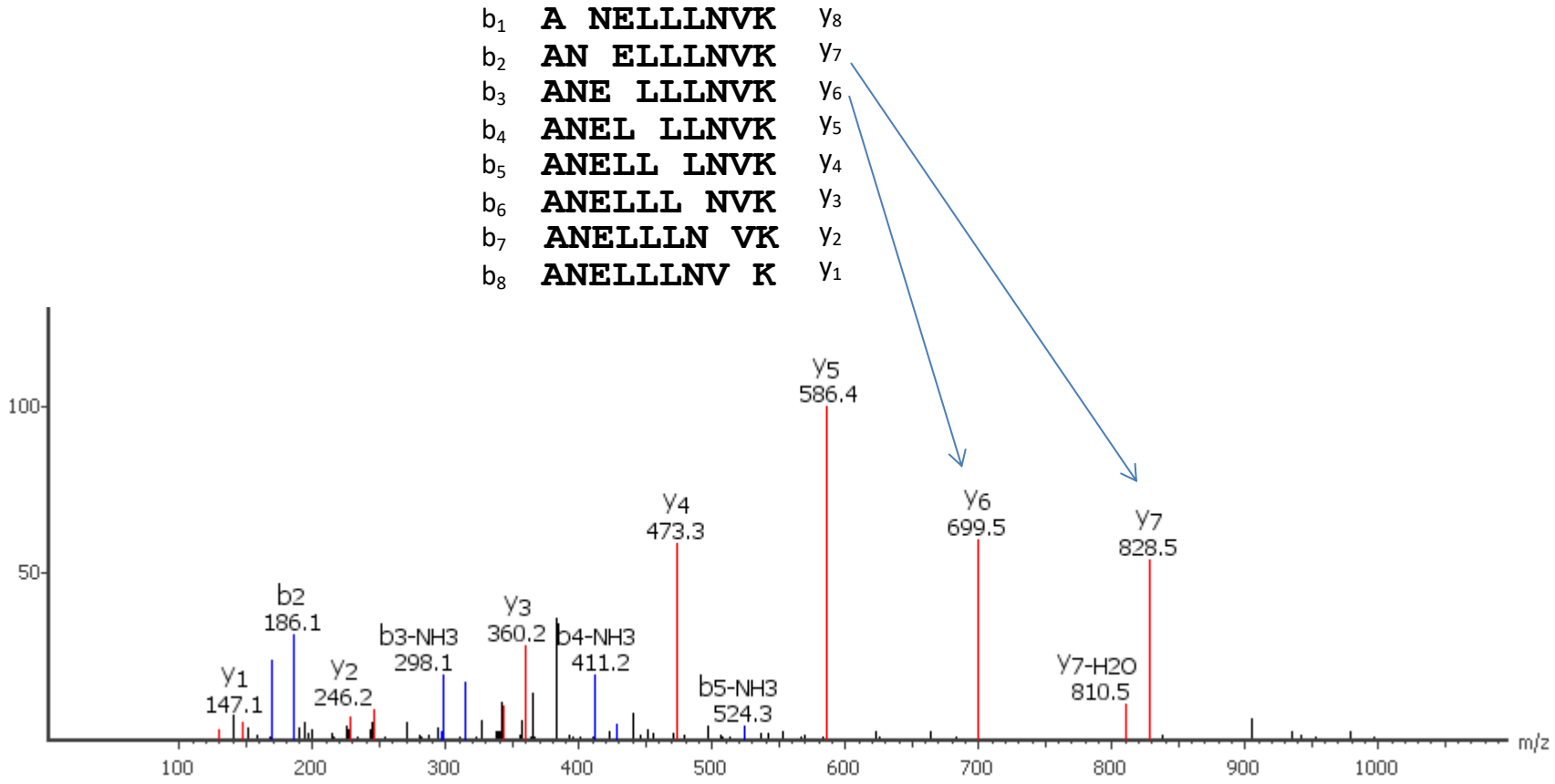


"Life is a series of tradeoffs, and greater results usually require greater tradeoffs."

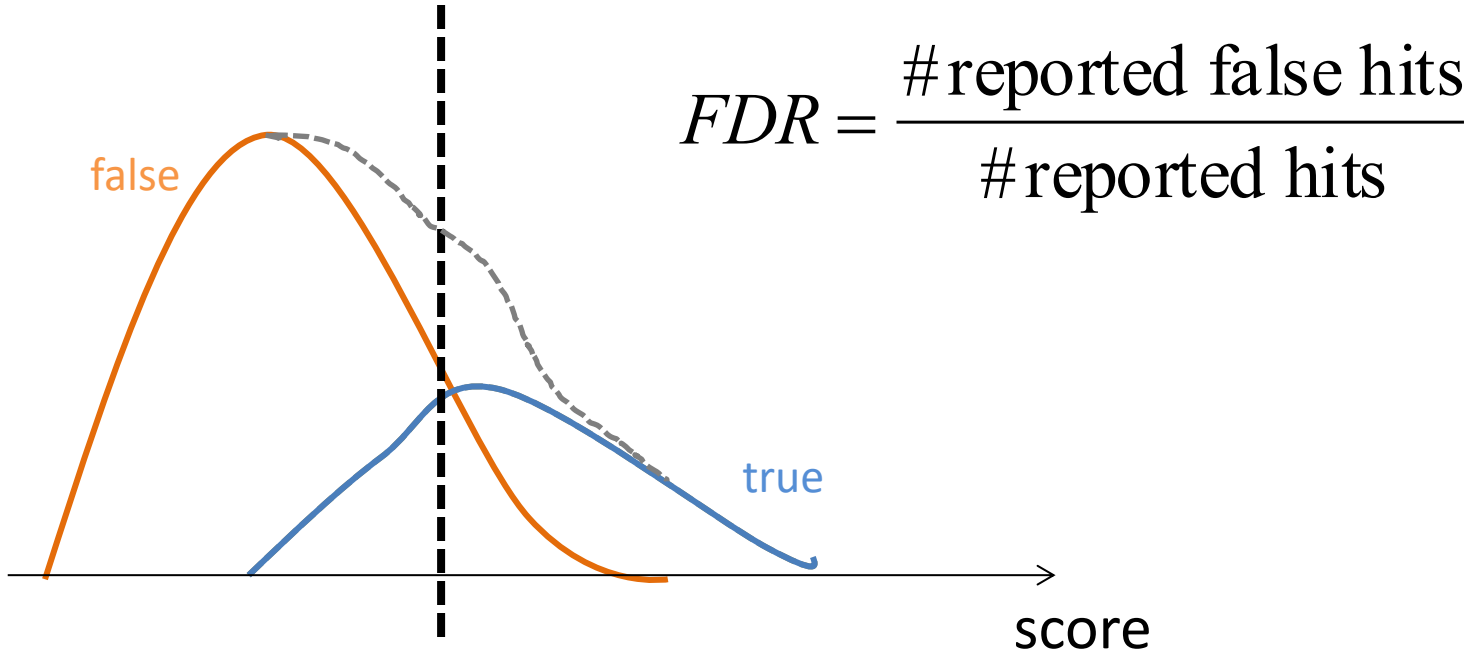
Bioinformatics Problems

- Seqeunce alignment and multiple sequence alignment
- Efficient homology search (seeding method)
- Mass spectrometry based proteomics
- Gene prediction
- NGS reads mapping (efficient index structure)
- Protein structure prediction

MS/MS of a Peptide



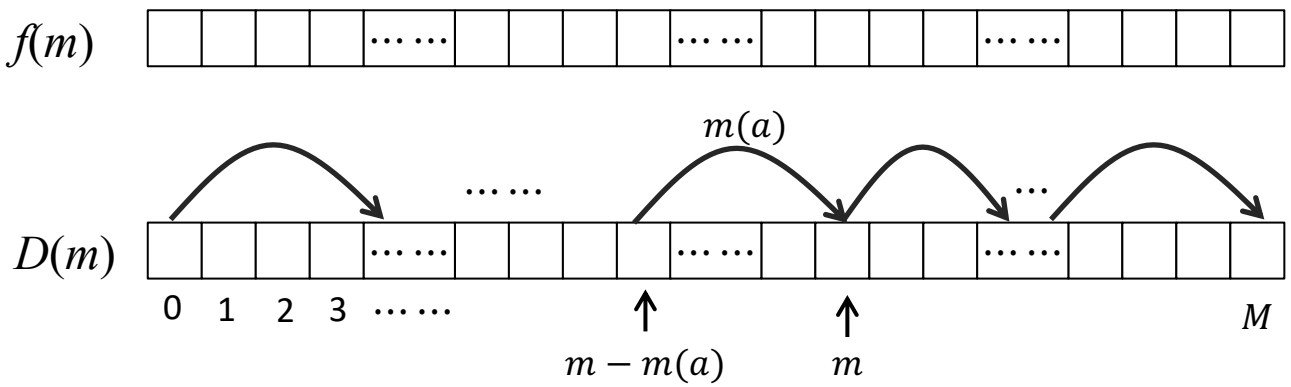
False Discovery Rate



$$FDR = \frac{\#\text{reported false hits}}{\#\text{reported hits}}$$

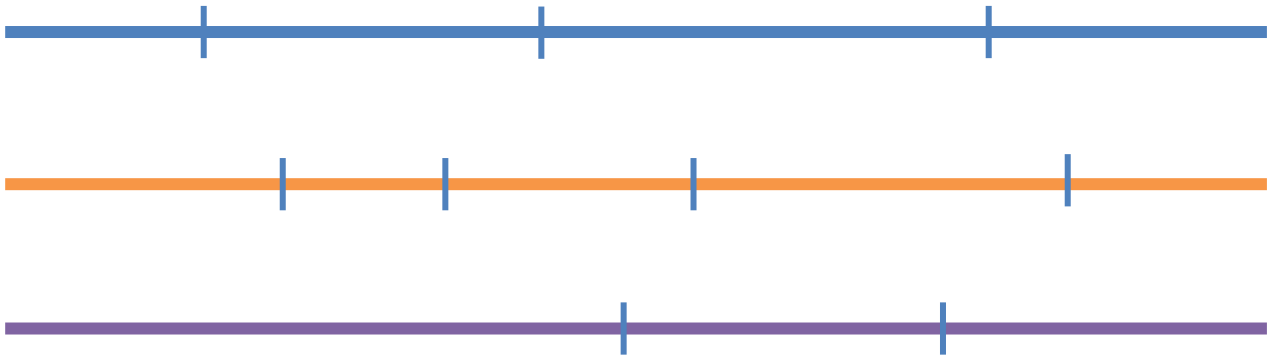
- By choosing different score threshold, one can calculate the FDR for all target PSMs above the threshold. Or conversely, one can choose a proper threshold to meet a FDR requirement.
- As of today, a typical FDR requirement is 1%.
- Unfortunately, we only know the aggregated distribution (grey curve)

De Novo Peptide Sequencing



- Let $D[m]$ be the maximum score a path from 0 to m can achieve.
- If the path is not empty, assume a is the last amino acid, then $D[m] = D[m - m(a)] + f(m)$.
- Thus, $D[m] = f(m) + \max_a D[m - m(a)]$.

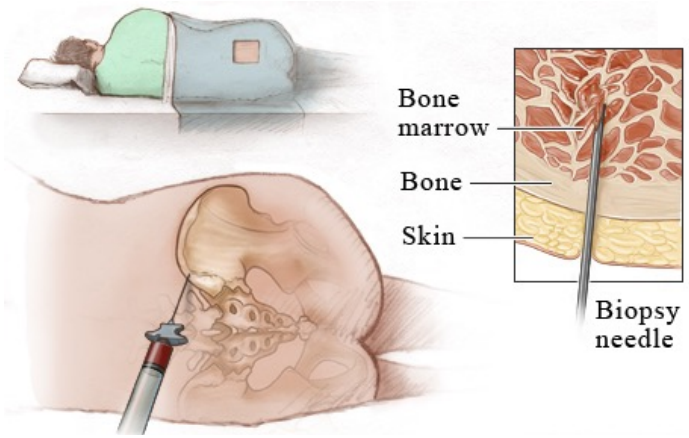
De Novo Protein Sequencing

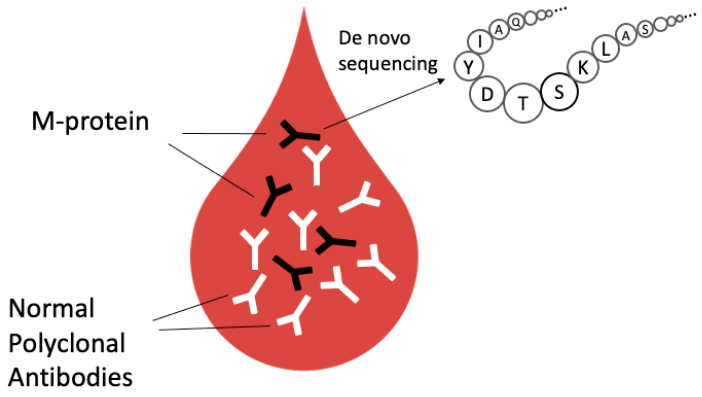


Trypsin digest

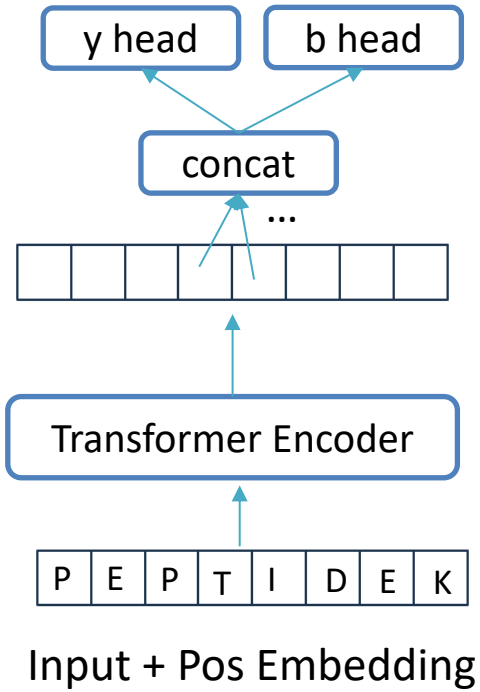
Chymotrypsin digest

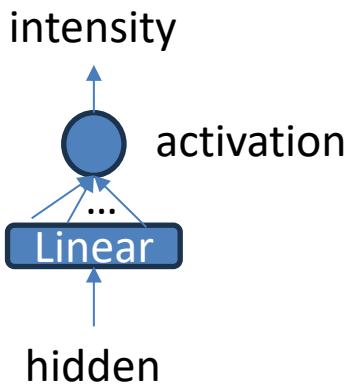
AspN digest





Spectrum Prediction





output head

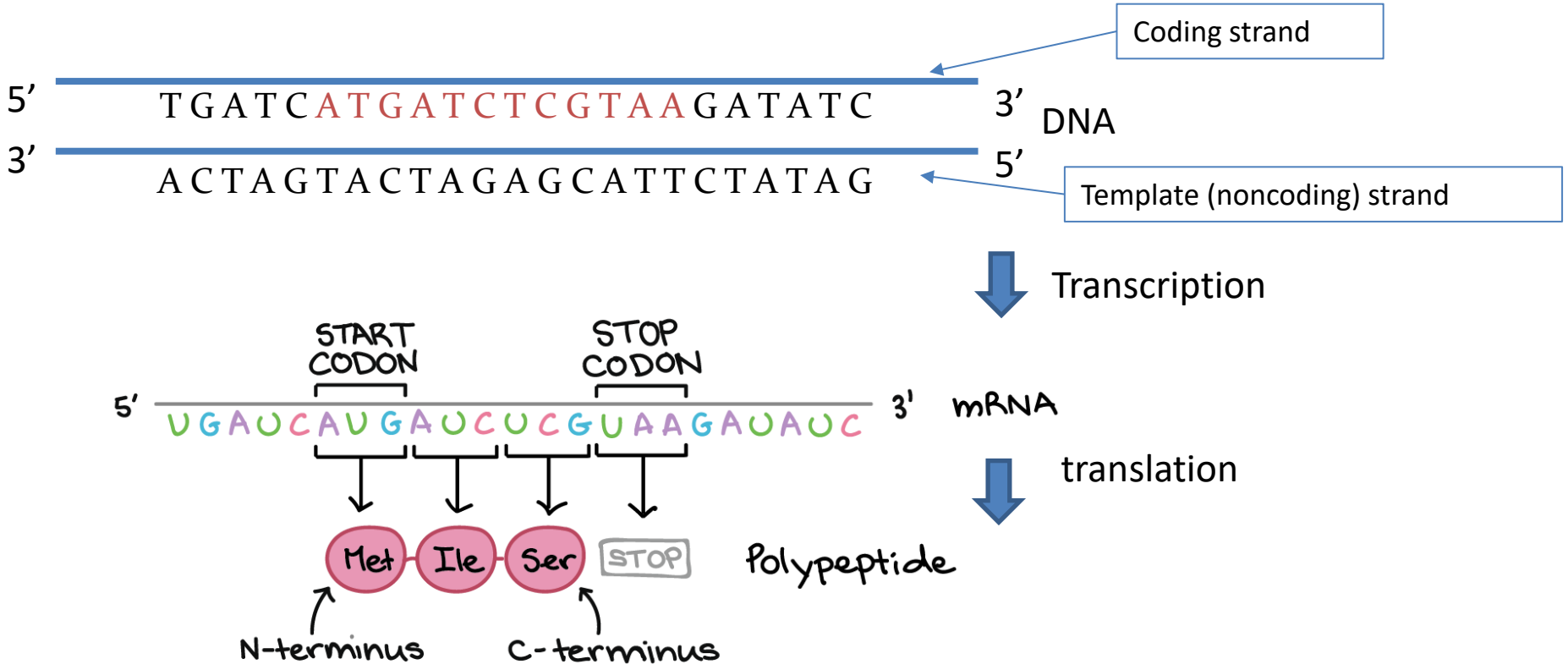
Label = Normalized intensity

Loss = MSE (Mean Square Error)

Bioinformatics Problems

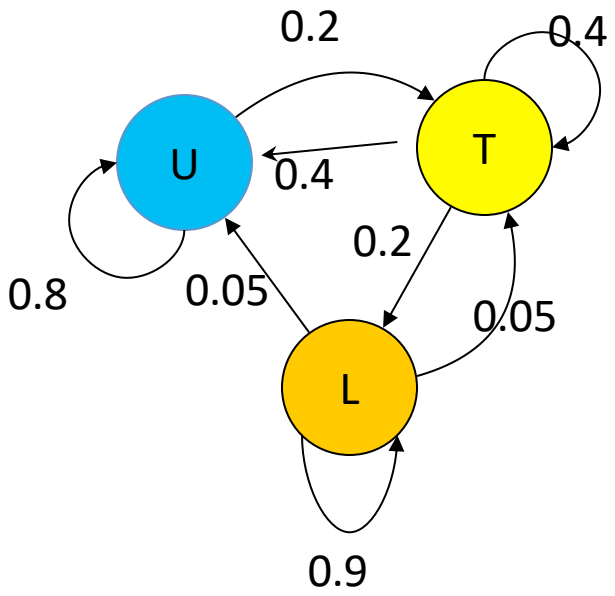
- Seqeunce alignment and multiple sequence alignment
- Efficient homology search (seeding method)
- Mass spectrometry based proteomics
- Gene prediction
- NGS reads mapping (efficient index structure)
- Protein structure prediction

From Gene to Protein (in *Prokaryotes*)



ORF: Open Reading Frame

Hidden Markov Model



U: Understands
T: Tries to understand
L: Lost completely

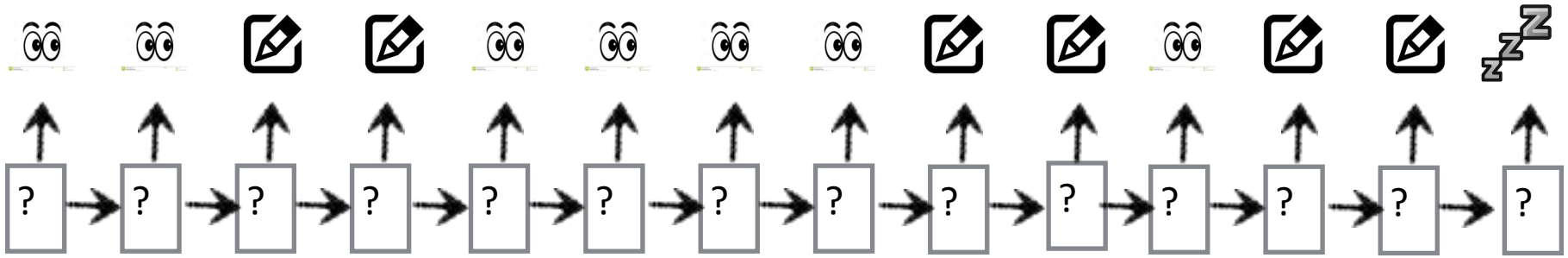
(T) Transition matrix

	U	T	L
U	0.8	0.2	0
T	0.4	0.4	0.2
L	0.05	0.05	0.9

(E) Emission matrix

	Look	Write	Sleep
U	0.6	0.35	0.05
T	0.9	0.1	0
L	0.1	0.6	0.3

Viterbi Algorithm



$$D[k,p] = \max_{p'} D[k-1,p'] \Pr(p|p') \Pr(S_i|p)$$

Gene Prediction Tool Tiberius

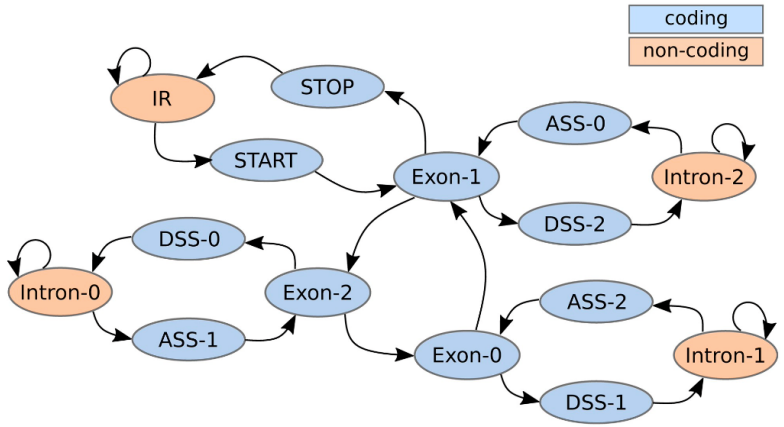


Figure 2. The states of the HMM used for inference with Tiberius and the transitions between them. The 11 coding-exon position states are subdivided by reading frame i: Exon-i represents non-border positions within an exon, while ASS-i (acceptor splice site) and DSS-i (donor splice site) states are the first and last position of an exon that starts and ends with reading frame *i*, respectively. The four non-coding position states are intergenic region (IR) or within an intron.

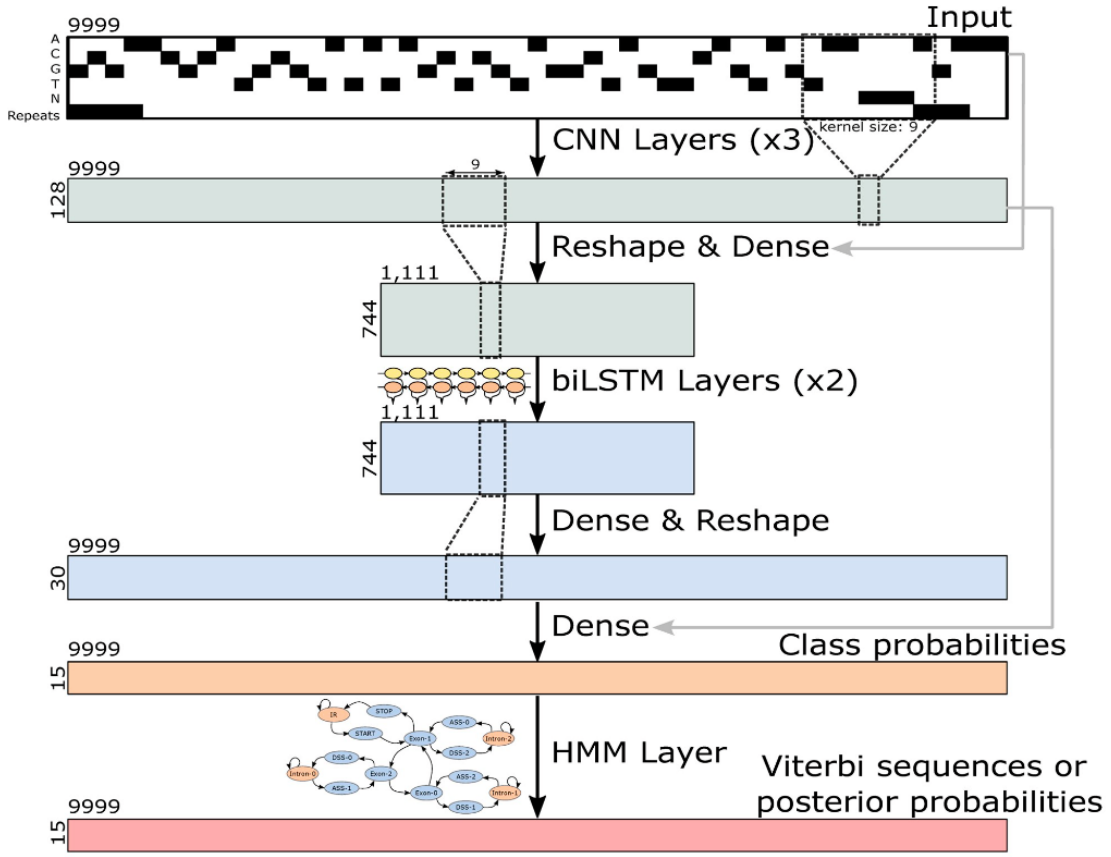


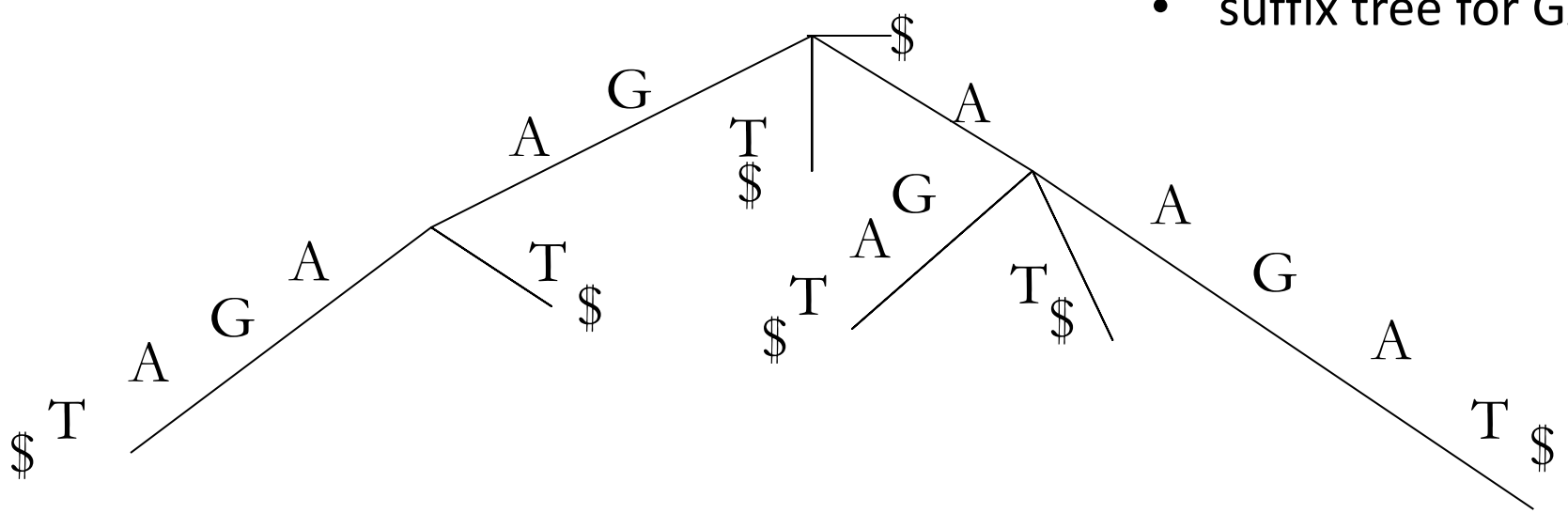
Figure 1. Illustration of the CNN-LSTM architecture of the Tiberius model for gene structure classification at each base position. The HMM layer computes posterior probabilities or complete gene structures (Viterbi sequences). The model has approximately 8 million trainable parameters, and it was trained with sequences of length $T = 9999$ and a length of $T = 500{,}004$ was used for inference.

Bioinformatics Problems

- Seqeunce alignment and multiple sequence alignment
- Efficient homology search (seeding method)
- Mass spectrometry based proteomics
- Gene prediction
- NGS reads mapping (efficient index structure)
- Protein structure prediction

Suffix Tree

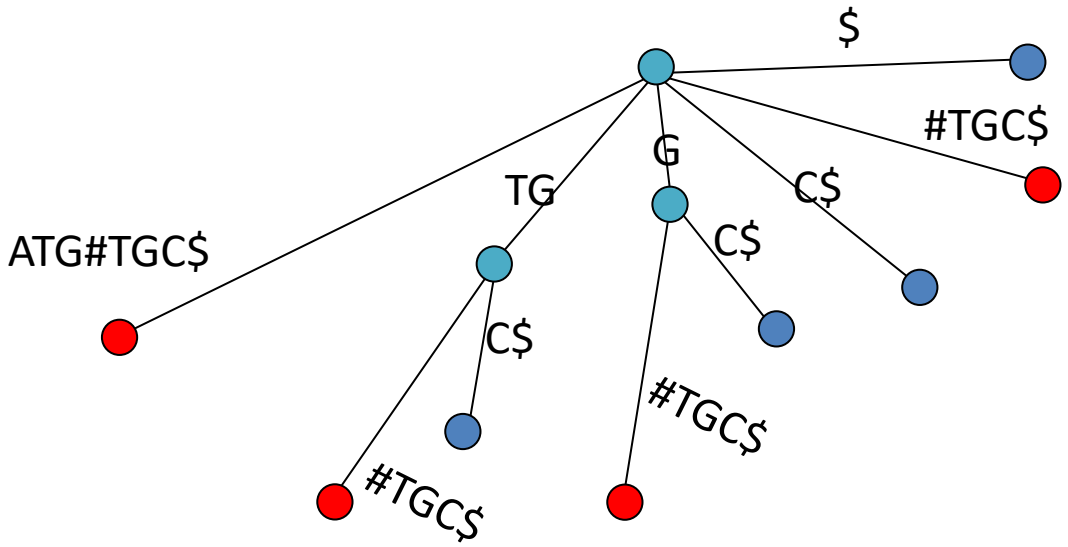
- suffix tree for GAAGAT$



- Pattern matching
- Longest common substring

Longest Common Substring

ATG#TGC$



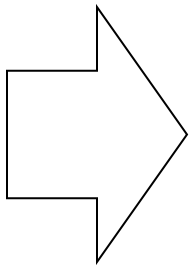
- Longest Common Substring
- Maximal Unique Matches

Suffix Array

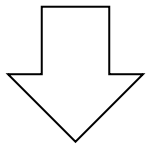
- Pattern matching
- Suffix sorting

- AGAAGAT$

1 = AGAAGAT$
2 = GAAGAT$
3 = AAGAT$
4 = AGAT$
5 = GAT$
6 = AT$
7 = T$
8 = $



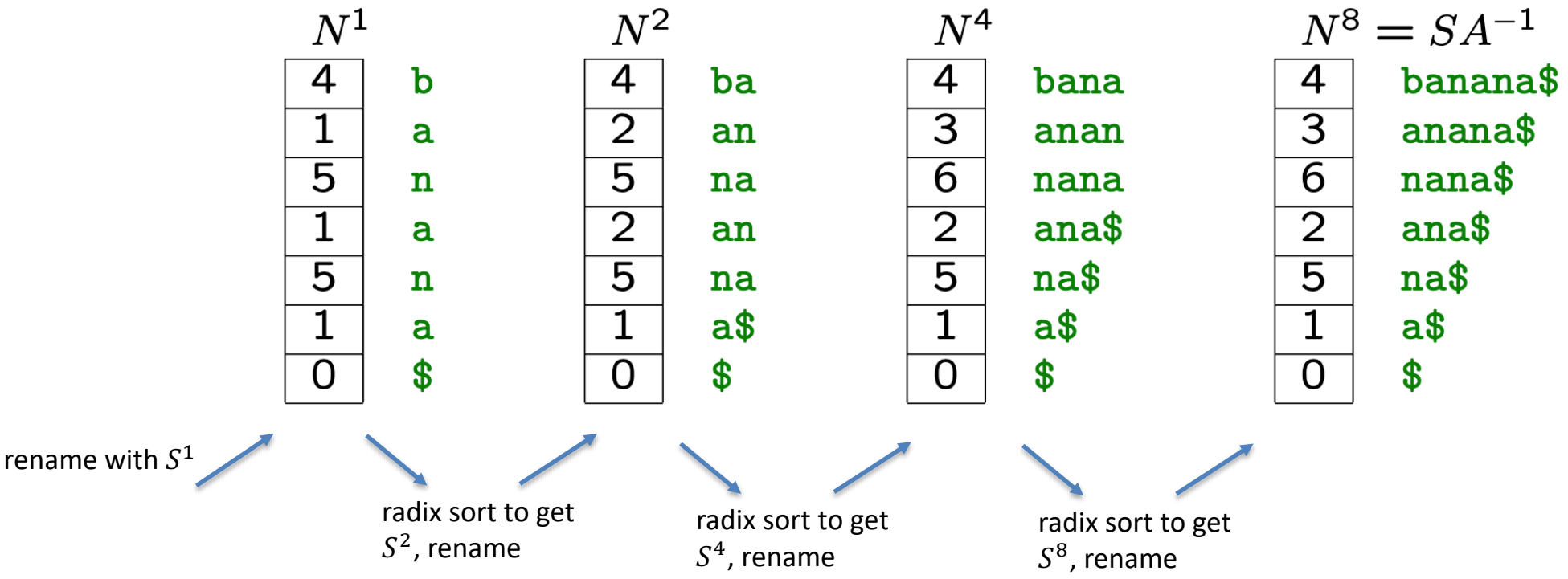
8 = $
3 = AAGAT$
1 = AGAAGAT$
4 = AGAT$
6 = AT$
2 = GAAGAT$
5 = GAT$
7 = T$



8, 3, 1, 4, 6, 2, 5, 7

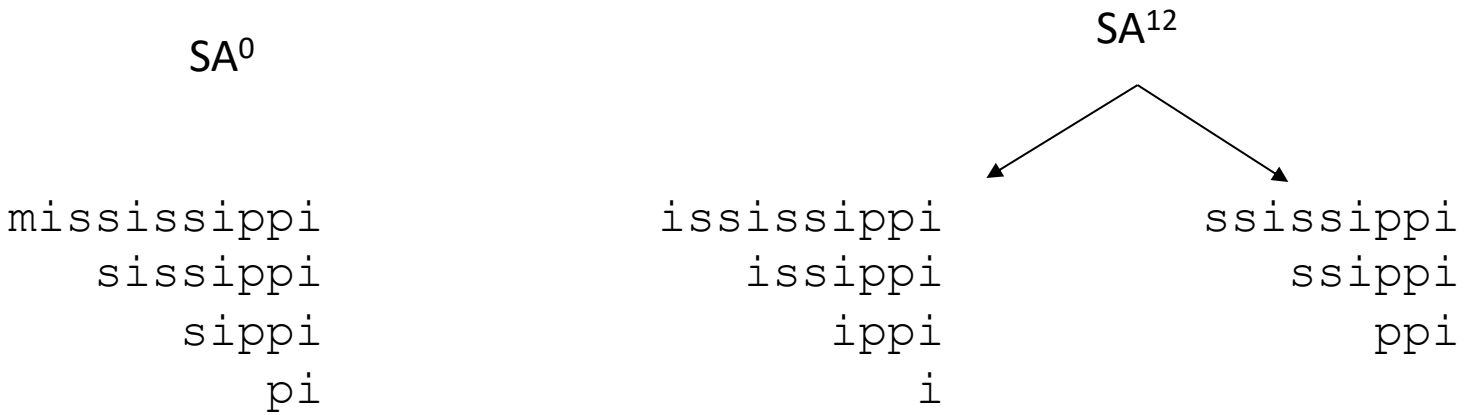
Prefix Doubling Algorithm

Renaming example for T=banana$

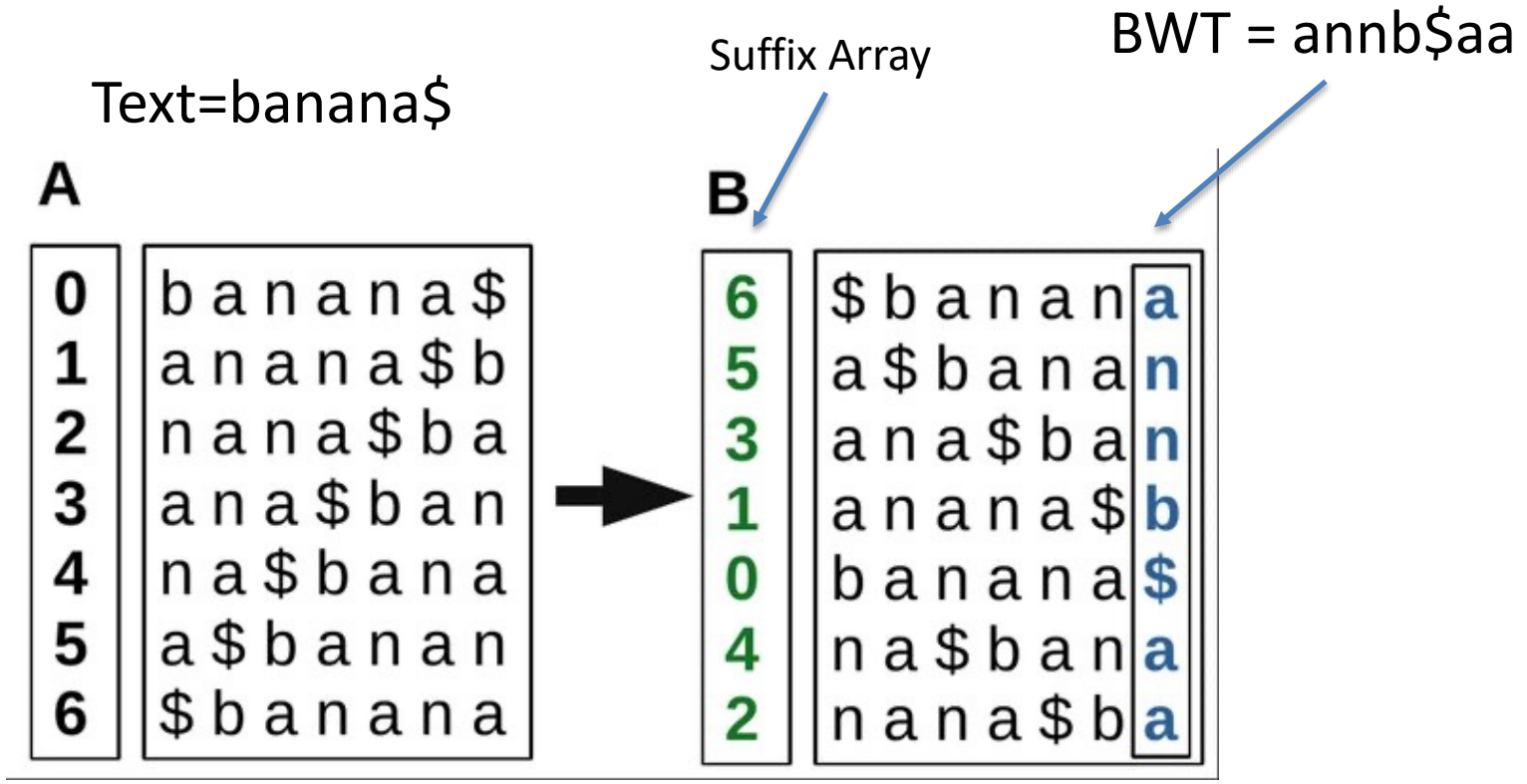


Skew Algorithm Example

- Example: mississippi

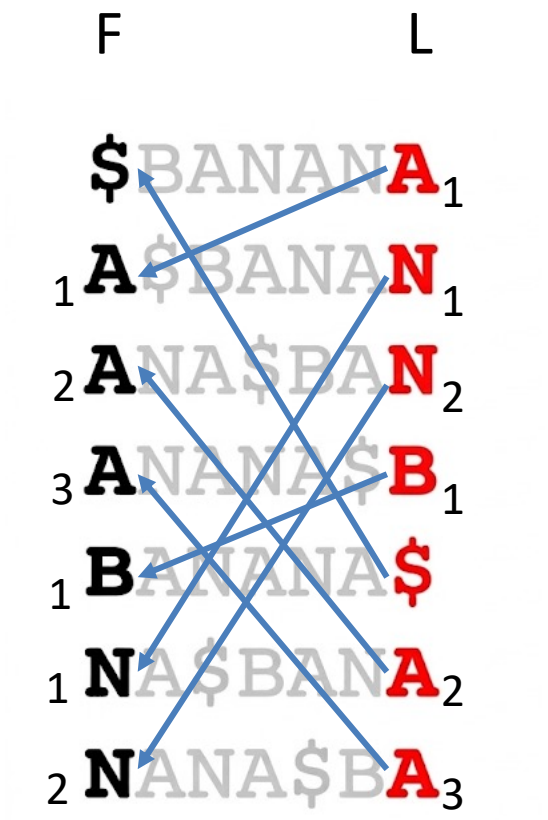


Burrows-Wheeler Transformation (BWT)



FM-Index

T= BANANA$



- LF-mapping
- Reconstruct T with LF-mapping
- Backward query with LF-mapping
- Store LF-mapping with C(a) and rank

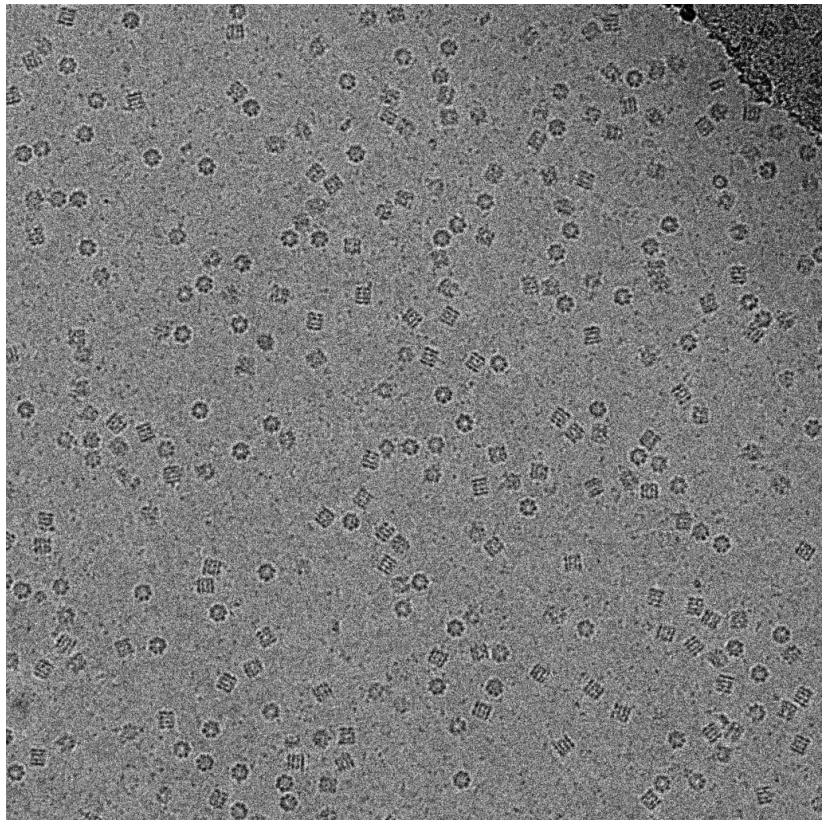
A Few General Skills

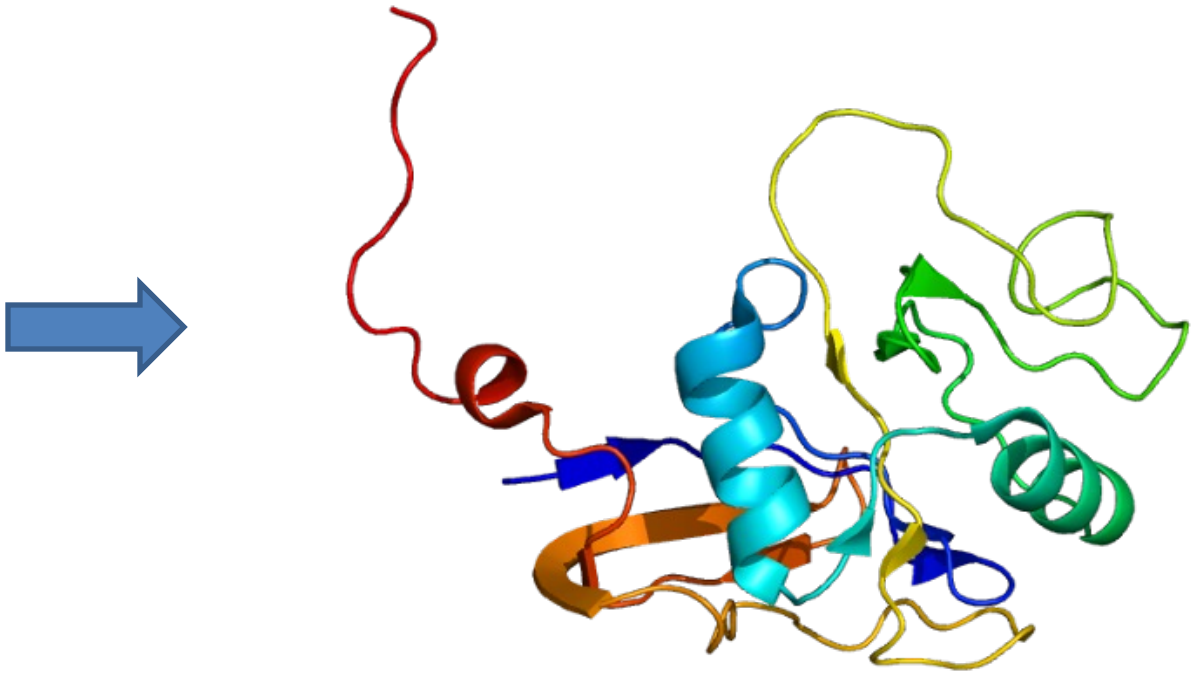
- Design a good scoring function
  - Log likelihood ratio
  - Evaluate the significance (Bayesian or p-value)
- Dealing with noisy data
  - FDR
- Dynamic programming
  - Sequence, set, mass …
- Trade off speed and accuracy/sensitivity
  - Filtration (Spaced seed)
- Useful models/data structure
  - HMM.
  - Suffix tree, suffix array, FM-Index
- Use of deep neural network

Other Things We Did Not Learn

- X-ray, NMR, Cryo EM for structure
- Single cell omics (genomics and proteomics)
- Other molecule types: Metabolomics, glycomics, RNA etc.
- Metagenomics/Microbiome
- De novo protein design

Cryo EM





Bioinformatics

- Experimental design
- Instrumentation
- Wet lab experimentas

Data

- Models
- Algorithms (including machine learning)
- Software

Molecules

Knowledge and solutions

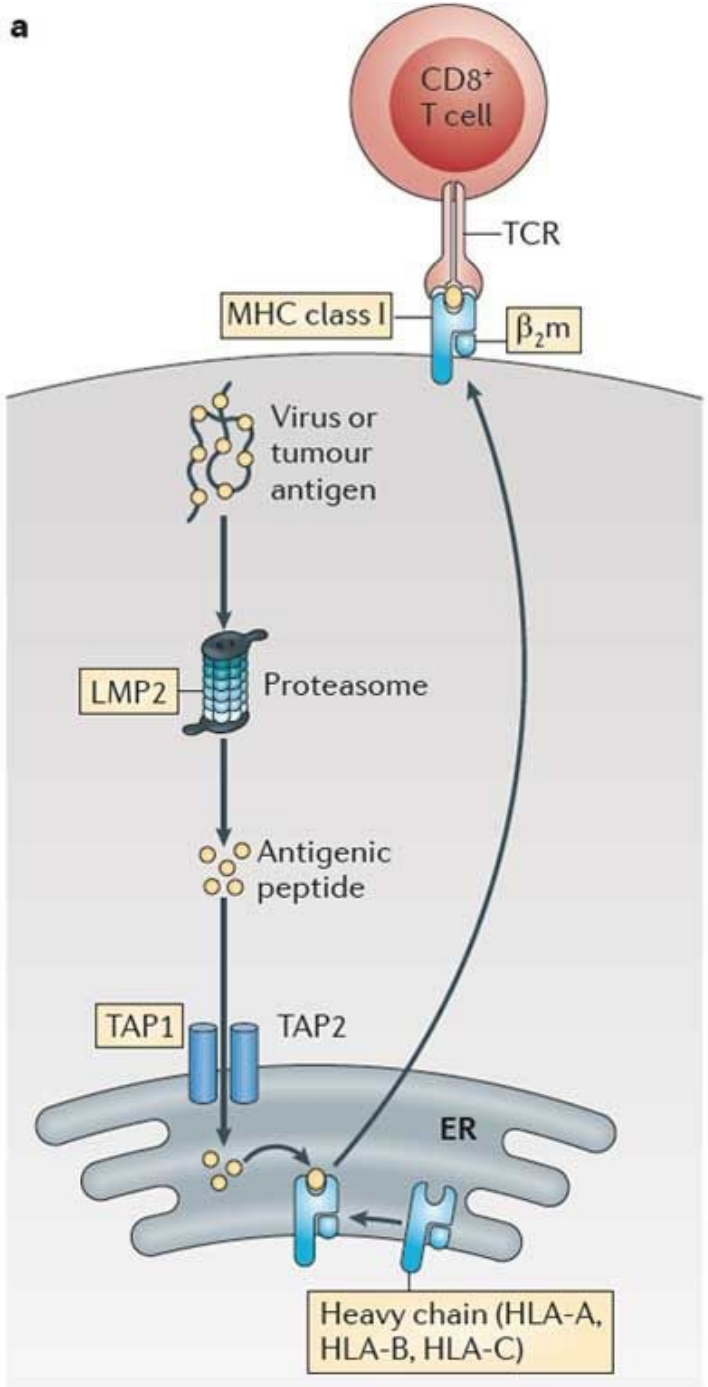
- Conjecture

Unmet needs, Unanswered questions

- Deployment
- Dissemination

Interesting News Article

- https://theconversation.com/a-man-used-ai-to-help-make-a-cancer-vaccine-for-his-dog-an-oncologist-urges-caution-278735



How Does It Work?

- Tumor or infected cells “present” some abnormal peptides at the MHC on cell surface.
- CD8+ T cells (aka T killer cells) recognize the abnormal peptides and kill the cell.
- Sequence the mRNA and predict the mutated peptides that can trigger T killer cell response.
- Those peptides may be too low to trigger the immune response.
- Synthesize mRNA that can produce the mutated peptides and use them as mRNA vaccine.
- Your body produces a lot of those peptides that enhance the T cell response to the target peptides, therefore kill the cancer cells.