Introduction to Genome Annotation Blast and HMMER methods

Adelaide Rhodes, Ph.D.
Global Invertebrate Genomics Alliance
October 20, 2018

- Phylum Bryozoa
- Aquatic Invertebrate
- “Moss Animals”
- Colonial

We are currently annotating the first genome in this phylum, *Bugula neritina*, host to a bacterial symbiont *Endobugula sertula* that biosynthesizes bryostatins, a class of anticancer molecules

Ernst Haeckl


2016 DEEPEND / Danté Fenolio

How do Marine Animals allow Symbionts to Colonize and Retain Immunity?

Common story in the marine realm: corals, squids, fish, etc. allow symbiotes to colonize

The immune system is altered to allow the hosted organism to thrive while repelling unrecognized invaders

It is not known how bryozoans have altered their immune systems to accommodate their symbionts. This is referred to as “innate immune recognition”.

Potential genes of interest: Pattern Recognition Receptors for Microbe-Associated Molecular Patterns (MAMPs)

Annotation Can help Answer this Question

Annotation is based on two strategies:

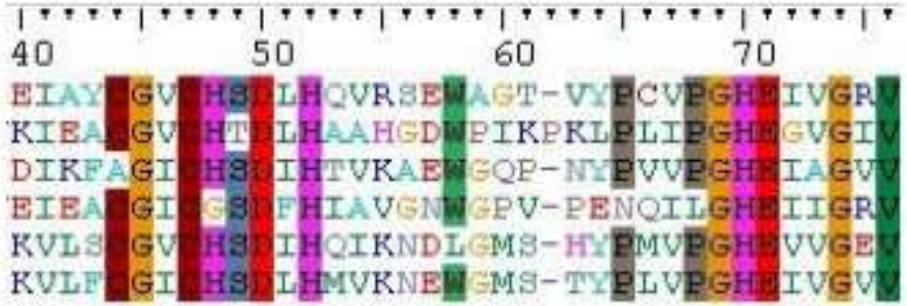
1.) Homology

2.) Ab initio prediction based on structure (Augustus, SNAP)

We are going to talk about homology today in order to answer the question of whether PRR found in other organisms are also present in the bryozoan *Bugula neritina*. We are also going to use two homology approaches to conduct whole genome annotation.

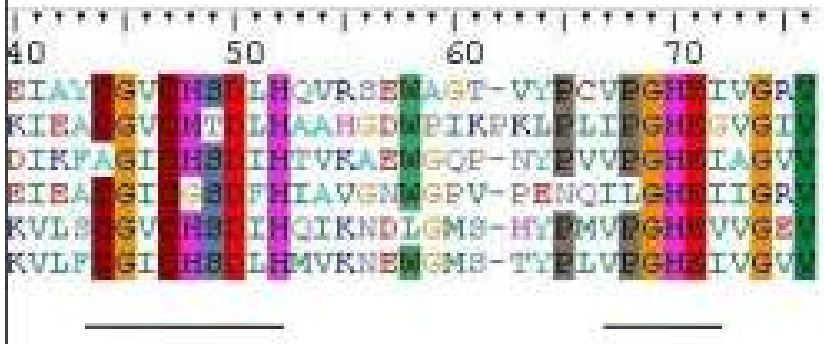
The basis for the prediction of features is nearly always a sequence alignment

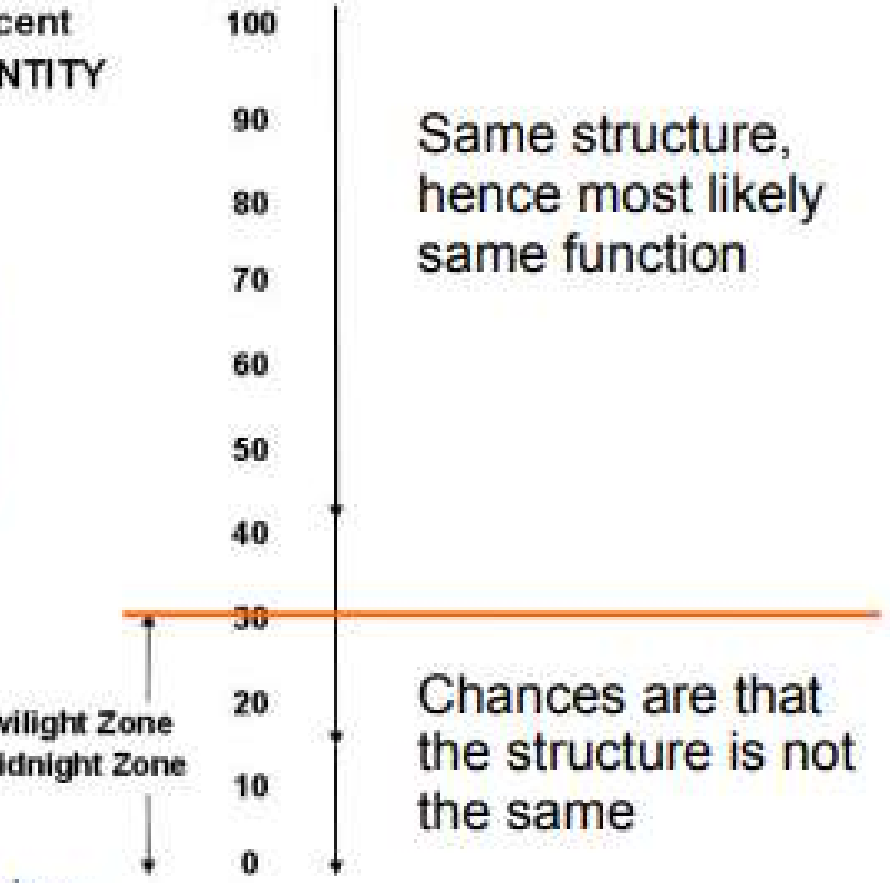
Based on experimentally verified sequence annotations, a multiple sequence alignment is constructed



Different methods exist to capture the information gained from this multiple sequence alignment

Alignment reveals similar residues which can indicate identical structure

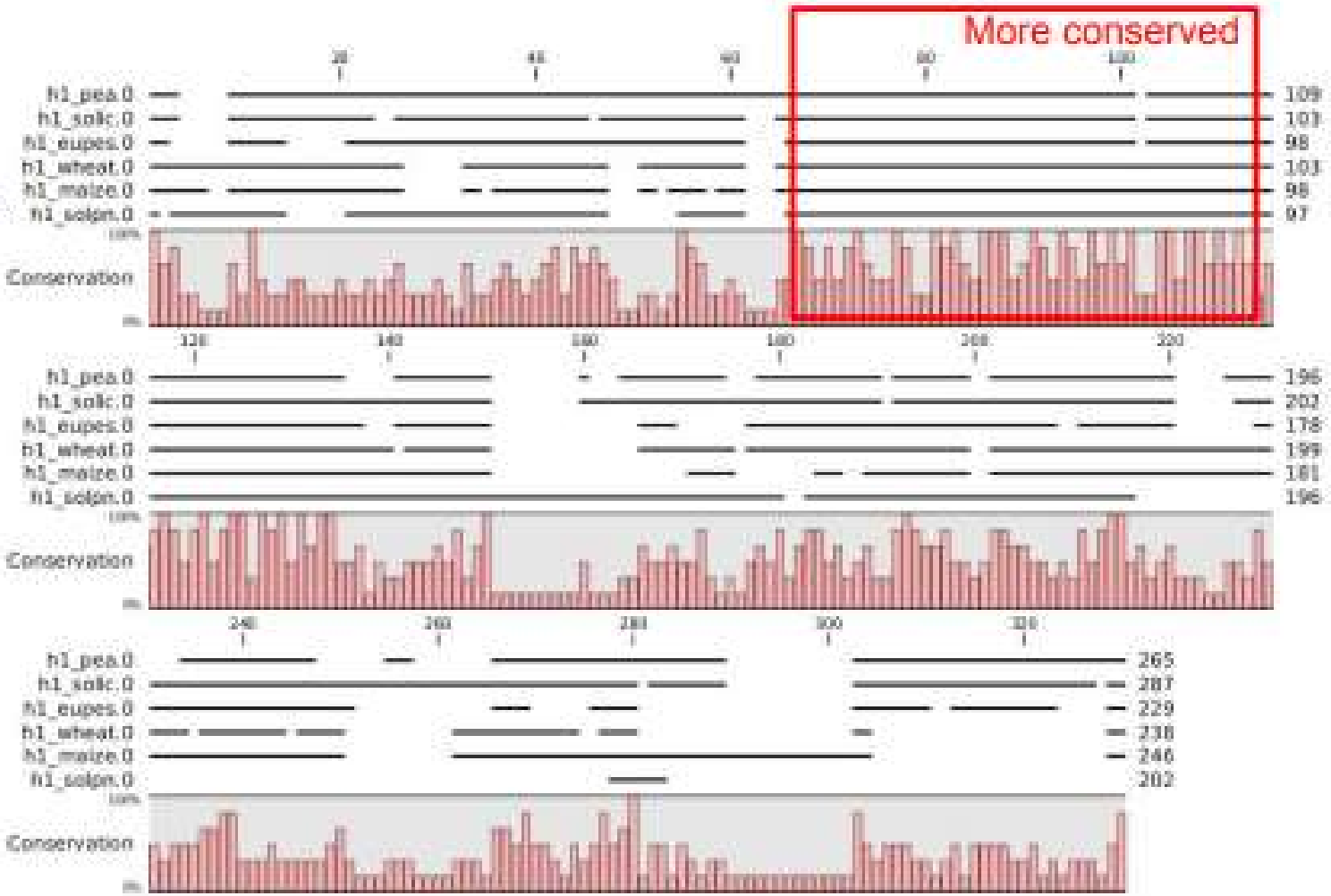




Most protein pairs with more than 25-30 out of 100 identical residues were found to be structurally similar. Also proteins with <10% identity can have similar structure.
http://peds.oxfordjournals.org/content/12/2/85.long

Degree of similarity with other sequences varies over the length

Homologous Histone H1 protein sequences



Protein sequences can consist of structurally different parts

Domain

part of the tertiary structure of a protein that can exist, function and evolve independently of the rest, linked to a certain biological function

Motif

part (not necessarily contiguous) of the primary structure of a protein that corresponds to the signature of a biological function. Can be associated with a domain.

Feature

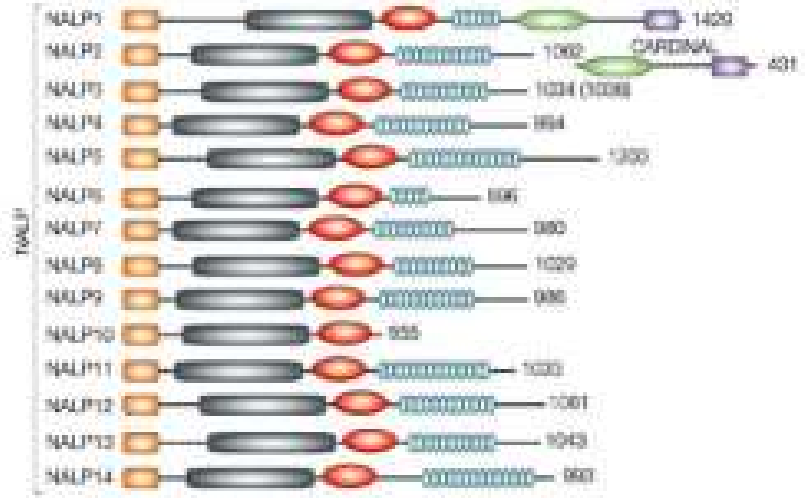
part of the sequence for which some annotation has been added. Some features correspond to domain or motif assignments.

Based on motifs and domains, proteins are assigned to families

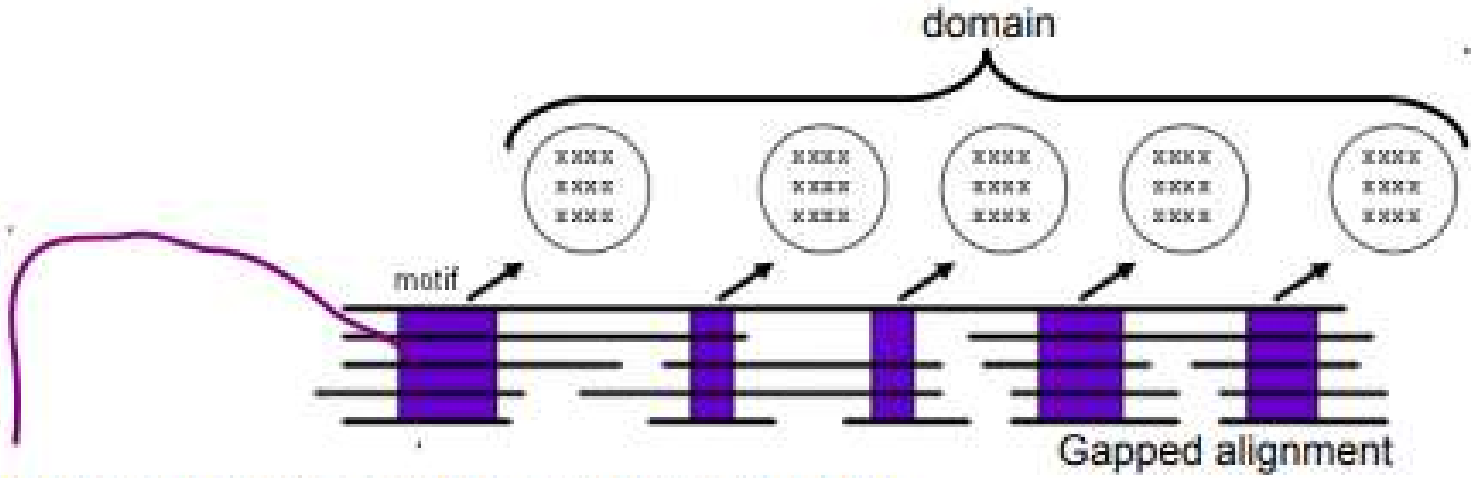
Nearly synonymous with gene family

Evolutionary related proteins

Significant structural similarity of domains is reflected in sequence similarity, and is due to a common ancestral sequence part, resulting in domain families.



Domains and motifs are represented by simple and complex methods



Motif/domain *in silico* can be represented by

1. Regular expression / pattern
2. Frequency matrix / profile
3. Machine learning techniques : Hidden Markov Model

http://bioinfo.uncc.edu/zhx/binf8312/lecture-7-SequenceAnalyses.pdf

How to Compare Two Sequences

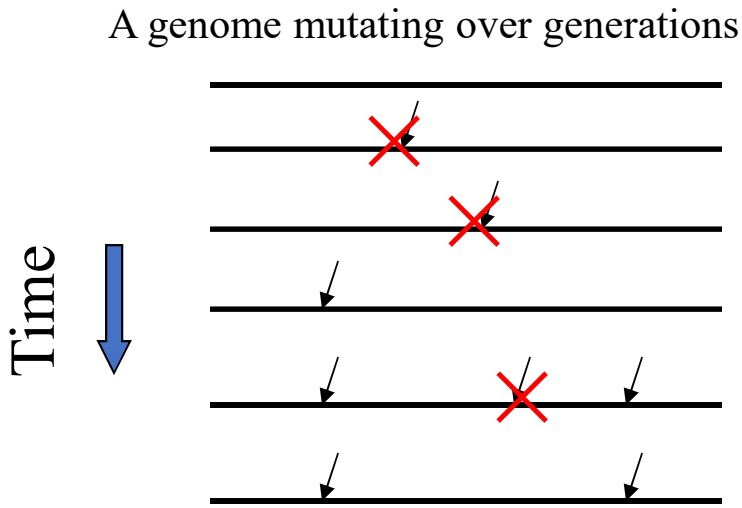
- **Problem**:
  - Given two sequences s_1 and s_2 over a fixed alphabet Σ, what is the set of variations that best describes the genetic transformation from s_1 to s_2 (or equivalently, from s_2 to s_1)?

Combinatorial Optimality	Probabilistic Optimality
• Based on either maximizing an *alignment score* or minimizing *edit distance*	• Based on finding a most *probable* set of changes in aligning two sequences
• Standard dynamic programming techniques (BLAST and MUSCLE)	• Hidden-Markov Model (HMM) techniques

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

Homology is important for Annotation

- *Mutation* ➔ natural genetic variations



- Mutations are random events
- The effect of only some mutation events carry over to future generations
- Sequence comparison key for evolutionary studies

substitution deletion insertion

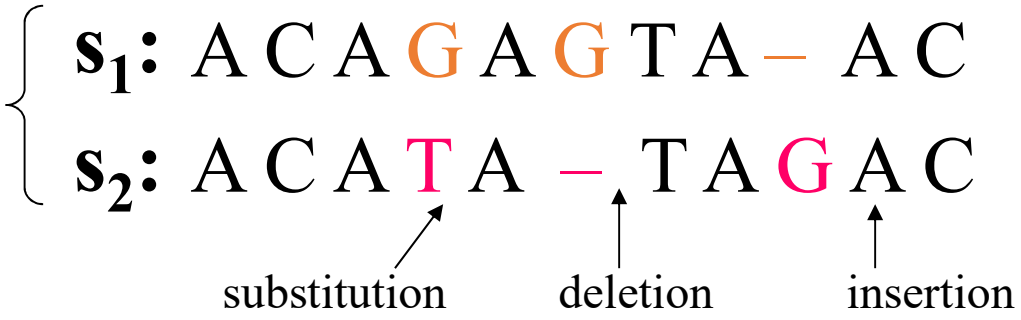
http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

Homology is important for Annotation

- *Mutation* ➔ natural genetic variations

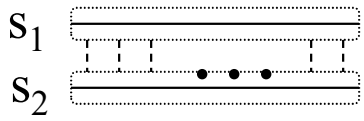
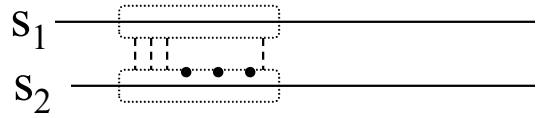
---

between s_1 and s_2



http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

Two Important Types of Alignments

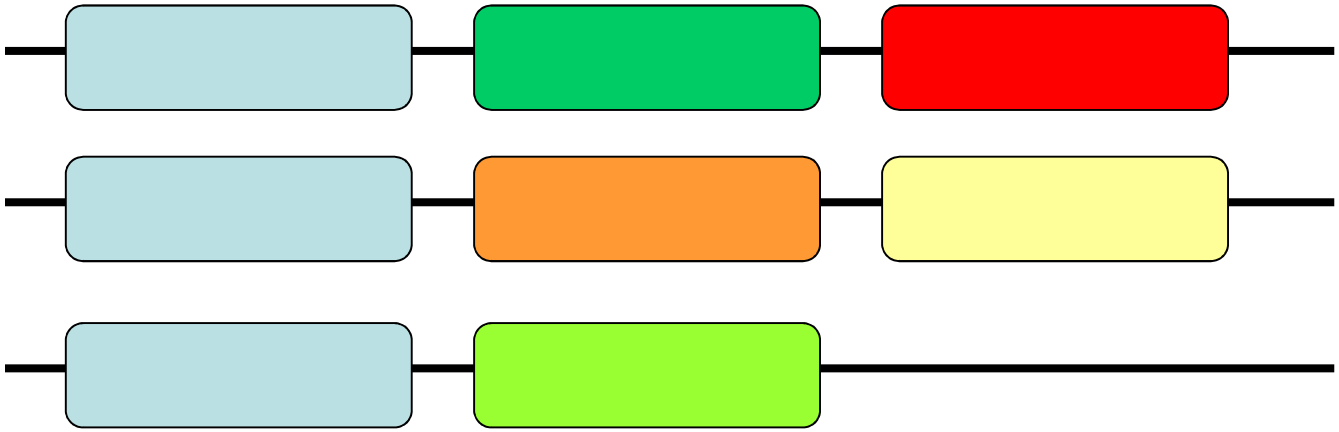
		Preferred Applications
Global Needleman-Wunsch	Alignment between s_1 and s_2  	For detecting two highly similar sequences (eg., two homologous proteins)
Local Smith-Waterman	Alignment between a substring of s_1 and a substring of s_2  	For detecting highly conserved regions (eg., genes) between two sequences (eg., genomes)

Optimal global and local alignments can be computed in $O(|s_1|.|s_2|)$ run-time and $O(|s_1|+|s_2|)$ space

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

Pairwise local/global alignment: differences

- Global alignment: we try to align the whole sequence. It is only useful for homologous proteins with a high percentage of identity.
- Local alignment: we try to align locally as much of the sequence as we can. This is useful when dealing with domains.



- Are these proteins homologues?
- Globally: no, they are very different, the score would be very low.
- Locally: there is a homologous domain, the grey one.

http://bioinfo.cnio.es/files/training/Fifth_Sequence_Analysis_2012/IntroductionToSequenceAnalysis2012Part2.ppt

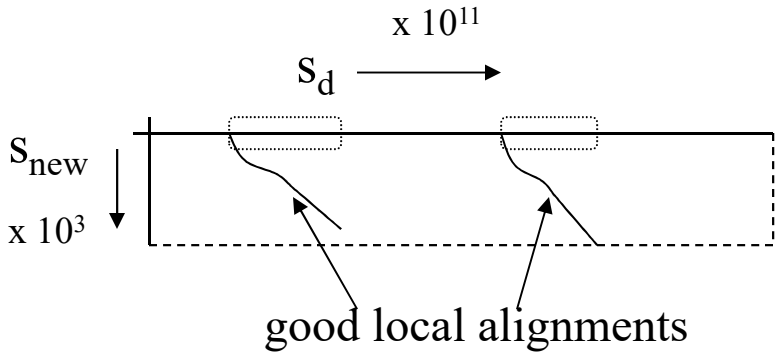
Need for a Fast Alignment Method

- What to do with a newly found gene candidate, s_{new}?
- Locate "similar" genes in GenBank

One-to-many

One Approach: (database search)

1. Concatenate all sequences in our genomic database into one sequence, say s_d
2. Compute the local alignment between s_{new} and s_d
3. Report all "significant" local alignments



Run-time: $O(|s_d|.|s_{new}|)$

⇩

Very long
query time !!

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

Basic Local Alignment Search Tool (BLAST)

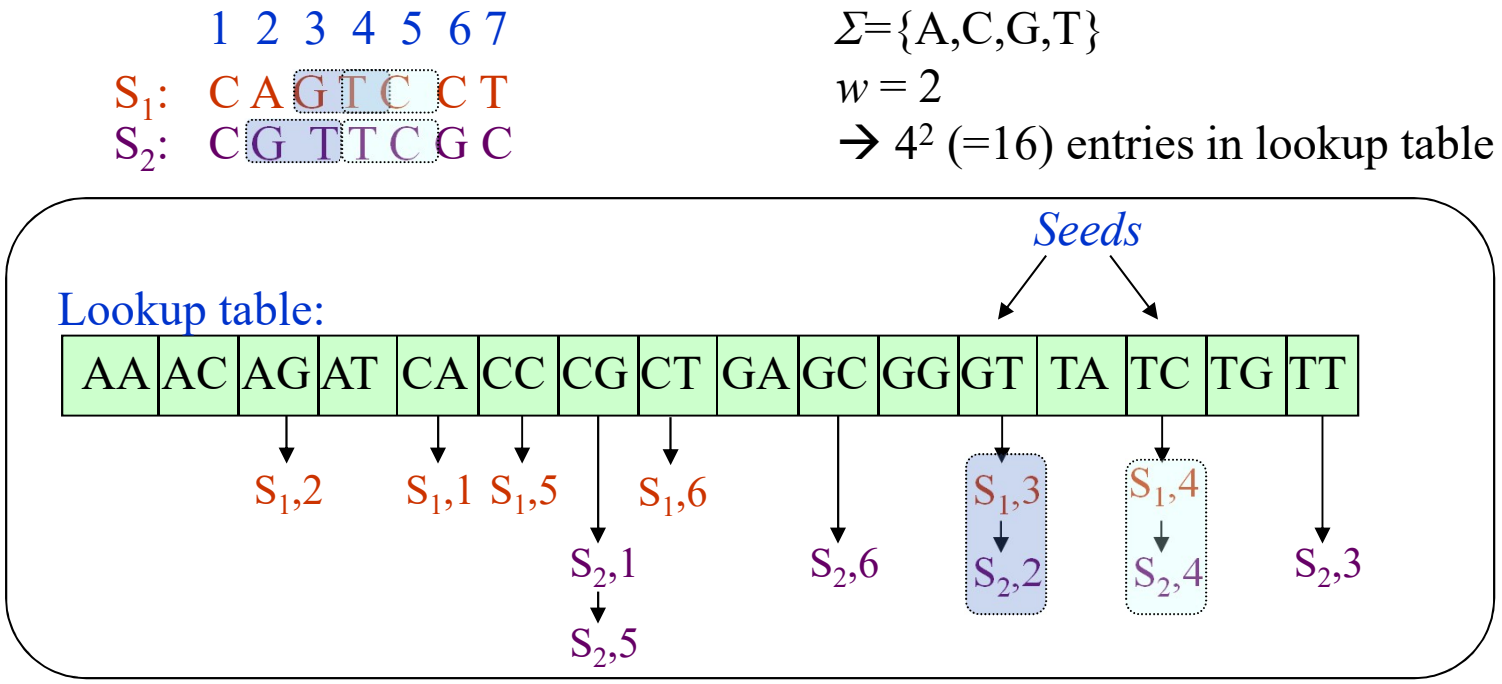
Altschul *et al.* (1990) developed a program called BLAST to quickly query large sequence databases

- **Input:**
  - Query sequence q and a sequence database D

- **Output:**
  - List of all significant local alignment hits ranked in increasing order of *E-value* (aka *p-value*, which is the probability that a random sequence scores more than q against D).

BLAST Algorithm

0. **Preprocess:** Build a *lookup table* of size $|\Sigma|^w$ for all w-length words in D



Preprocessing is a one time activity

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

BLAST Algorithm …

Identify Seeds*:* Find all *w*-length substrings in *q* that are also in D using the lookup table

Extend seeds: Extend each seed on either side until the aggregate alignment score falls below a threshold

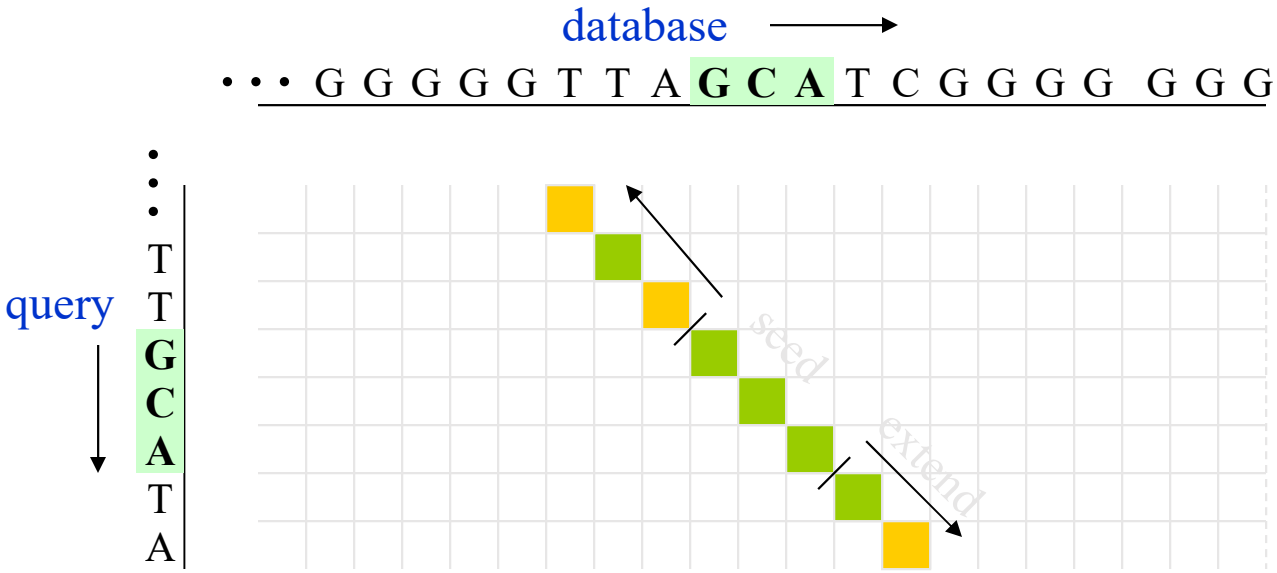
- Ungapped: Extend by only either matches or mismatches
- Gapped: Extend by matches, mismatches or a limited number of insertion/deletion gaps

Record all local alignments that score more than a certain statistical threshold

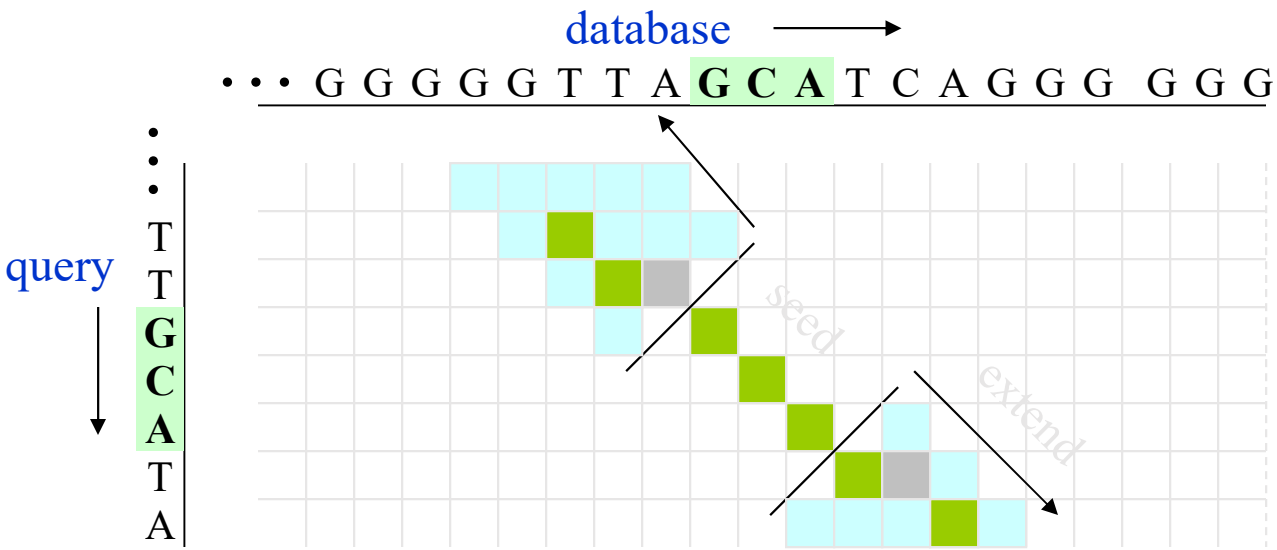
Rank and report all local alignments in non-decreasing order of *E-value*

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

Illustration of BLAST Algorithm



Ungapped Extension



Gapped Extension (over a band of diagonals)

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.pp

Different Types of BLAST Programs

Program	Query	Database
blastn	nucleotide	nucleotide
blastp	protein/peptide	protein/peptide
blastx	nucleotide	protein/peptide
tblastn	protein/peptide	nucleotide
tblastx	nucleotide	nucleotide

http://www.ncbi.nlm.nih.gov/blast

http://www.shodor.org/media/content/petascale/materials/dataIntensive/BLAST/BLAST_Intro_ppt.ppt

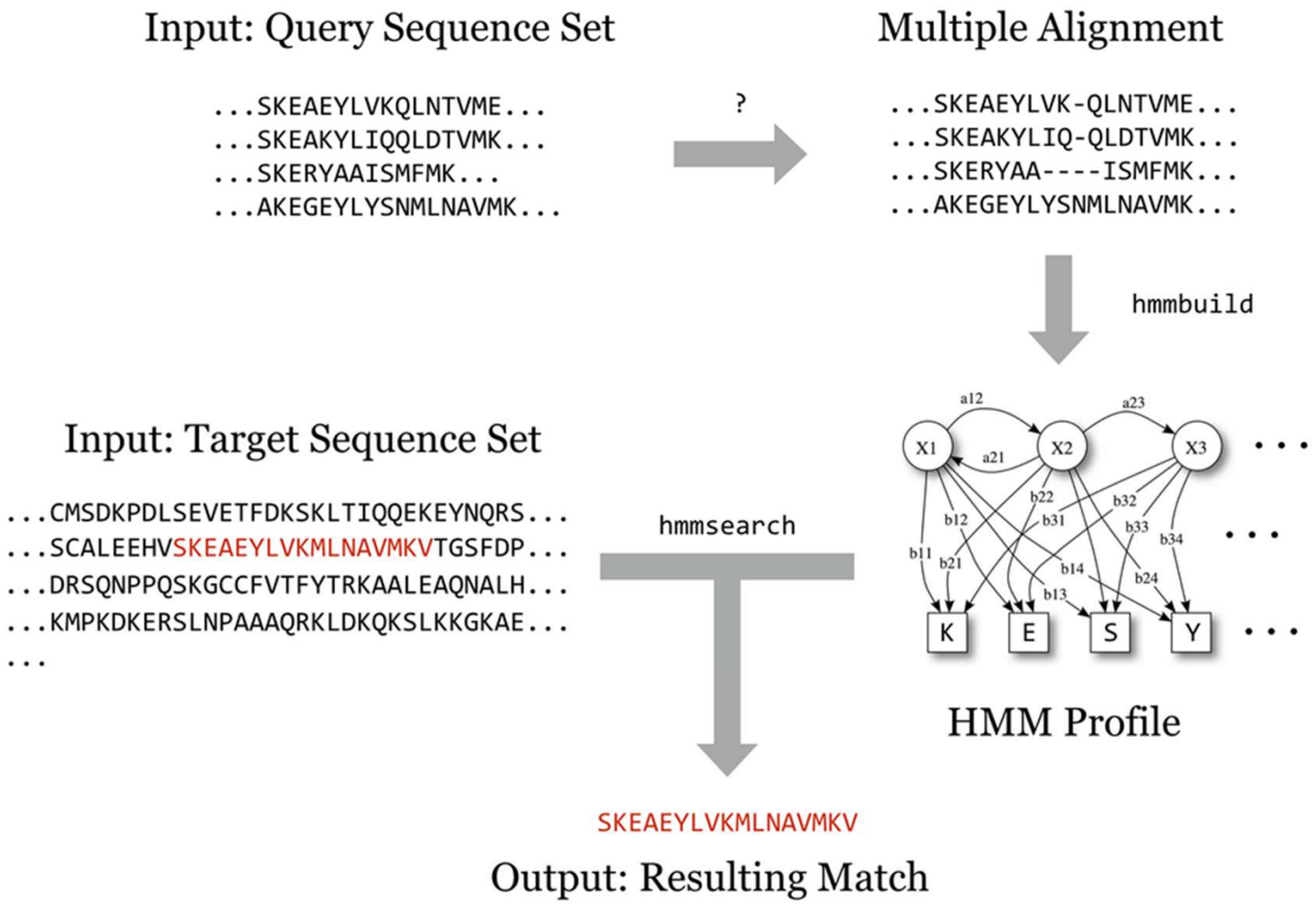
Global Alignments and PROBABILISTIC APPROACHES to Homology

- Because Blast is computationally greedy, many annotation programs are turning to a probabilistic approach.
- HMMER – Builds a profile based on a training data set.
- Similar to other “machine learning approaches.
- More Sensitive [we will test this in the tutorial]

HMM starts with a Multiple Sequence Alignment

```
CLUSTAL 2.0.12 multiple sequence alignment


sp|O88479|FOS_MESAU       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLSVSSANF 60
sp|Q56TN0|FOS_PHORO       MMFSGFNTDYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCADLSVSSANF 60
sp|O77628|FOS_BOVIN       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDYCTDLAVSSANF 60
sp|Q8HZP6|FOS_FELCA       MMFSGFNADYEASSSRCSSASPAGDNLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF 60
sp|P01100|FOS_HUMAN       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF 60
sp|P12841|FOS_RAT         MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
sp|P01102|FOS_MSVFB       MMFSGFNADYEASSFRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
sp|P11939|FOS_CHICK       MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
sp|P53539|FOSB_HUMAN      -MFQAFPGDYDSGS-RCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
sp|Q9TUB3|FOSB_CANFA      -MFQAFPGDYDSGS-RCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
sp|P13346|FOSB_MOUSE      -MFQAFPGDYDSGS-RCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
                           *:.:*  :*:: * **** **:.:.   * *..***.*  :.. :*: *:.*.  ...*


sp|O88479|FOS_MESAU       IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPTPS-----TGAYSR 108
sp|Q56TN0|FOS_PHORO       IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPTPS-----TGAYSR 108
sp|O77628|FOS_BOVIN       IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPTPS-----AGAYSR 108
sp|Q8HZP6|FOS_FELCA       IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPAPS-----AGAYSR 108
sp|P01100|FOS_HUMAN       IPTVTAISTSPDLQWLVQPALVSSVAPS-------QTRAPHPFGVPAPS-----AGAYSR 108
sp|P12841|FOS_RAT         IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGLPTPS-----TGAYAR 108
sp|P01102|FOS_MSVFB       IPTVTATSTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGLPTQS-----AGAYAR 108
sp|P11939|FOS_CHICK       VPTVTAISTSPDLQWLVQPTLISSVAPS-------QNRG-HPYGVPAPAP----PAAYSR 108
sp|P53539|FOSB_HUMAN      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSS 114
sp|Q9TUB3|FOSB_CANFA      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGMSGYSS 114
sp|P13346|FOSB_MOUSE      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYST 114
                          :***** :** ********:*:**:* *       *    .*:.:*  :      ..*:


sp|O88479|FOS_MESAU       --------AGMVKTVSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|Q56TN0|FOS_PHORO       --------AGMVKTVSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|O77628|FOS_BOVIN       --------AGVMKTMTGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|Q8HZP6|FOS_FELCA       --------AGVVKTVTAGG---RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 157
sp|P01100|FOS_HUMAN       --------AGVVKTMTGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|P12841|FOS_RAT         --------AGVVKTMSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|P01102|FOS_MSVFB       --------AEMVKTVSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|P11939|FOS_CHICK       --------PAVLK-APGG----RGQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 155
sp|P53539|FOSB_HUMAN      GGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
sp|Q9TUB3|FOSB_CANFA      GGASGSGGPSTSGTTSGPGPARPARARLRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
sp|P13346|FOSB_MOUSE      GGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
                                  .      ..      .::  ** : * *;********;******;*******


sp|O88479|FOS_MESAU       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
sp|Q56TN0|FOS_PHORO       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEDM 216
sp|O77628|FOS_BOVIN       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
sp|Q8HZP6|FOS_FELCA       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 217
sp|P01100|FOS_HUMAN       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
```

HMMER from MSA



Frequency matrices or profiles include the chance of observing the residues

For every position of a motif, a list of all amino acids is made with their frequency. Position-specific weight/scoring matrix or profile. More sensitive way.

Profile

	123456
	ATPKAE
	KKPKAA
	AKPKAK
	TKPKPA
	AKPKT-
	AKPAAK
	KLPKAD
	AKPKAA
Consensus:	AKPKA-

Position:	1.	2.	3.	4.	5.	6.
A	0.625	0	0	1/8	6/8	3/8
D	0	0	0	0	0	1/8
E	0	0	0	0	0	1/8
K	0.25	6/8	0	7/8	0	2/8
L	0	1/8	0	0	0	0
P	0	0	1	0	1/8	0
T	1/8	1/8	0	0	1/8	0
-	0	0	0	0	0	1/8
Sum	1	1	1	1	1	1

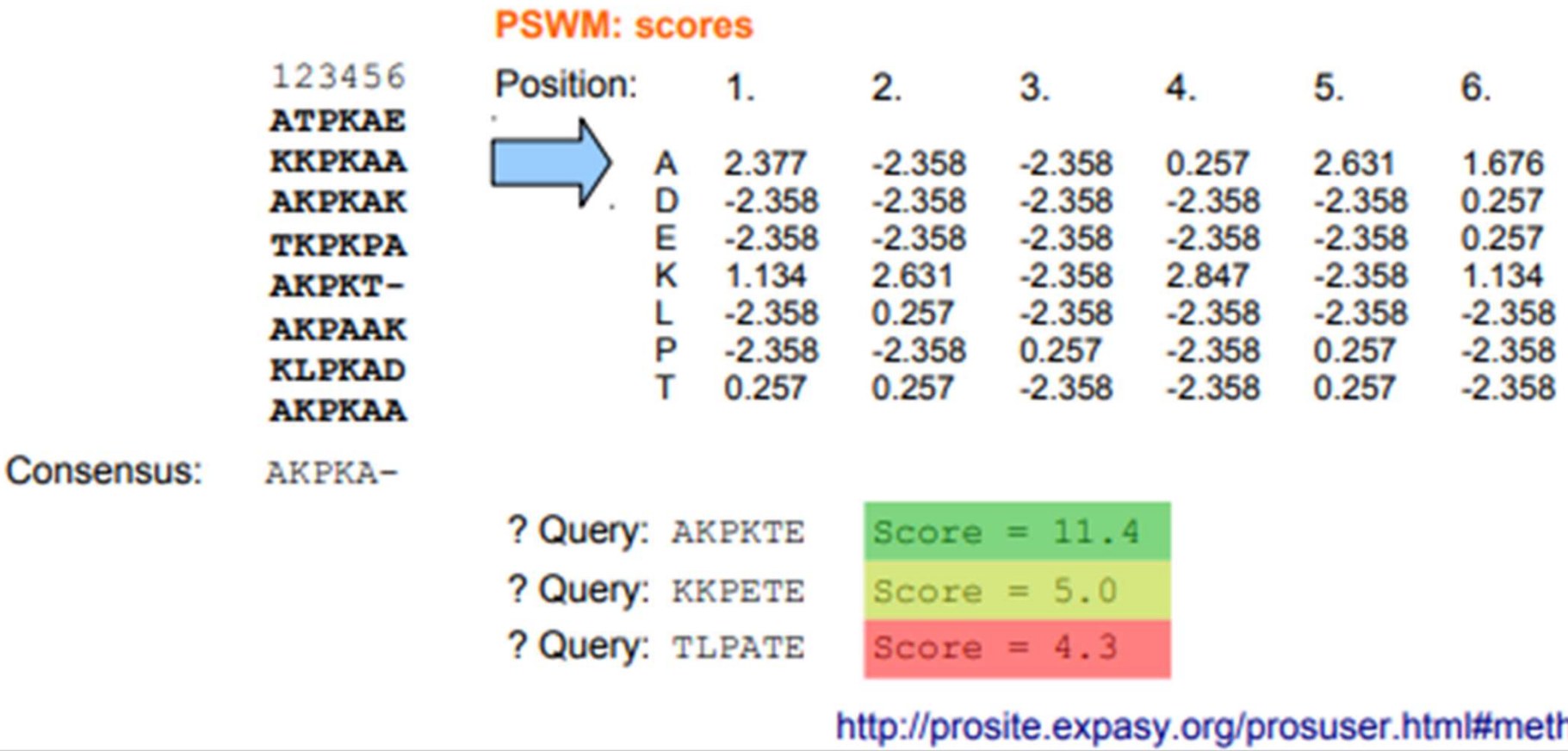
? Query: AKPKTE

? Query: KKPETE

? Query: TLPATE

Example: http://expasy.org/prosite/PS51092 http://prosite.expasy.org/prosuser.html#meth2

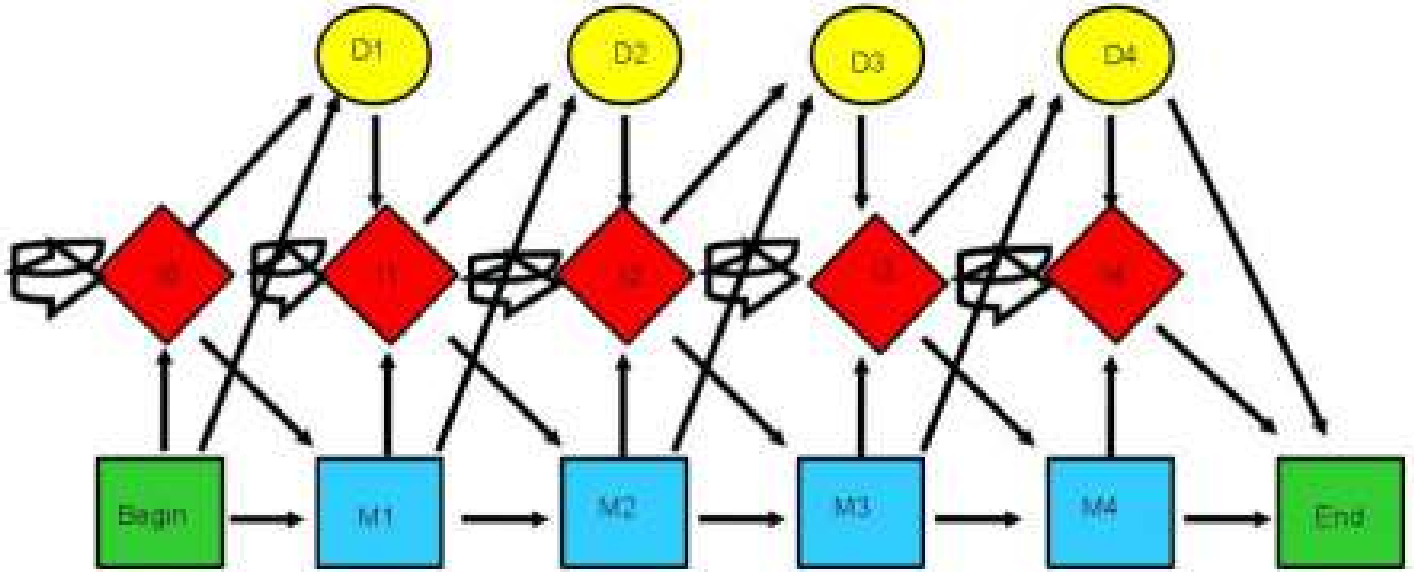
How good a sequence matches a profile is reported with a score



http://prosite.expasy.org/prosuser.html#meth2

A hidden Markov Model takes also into account the gaps in an alignment

The schematic representation of a HMM

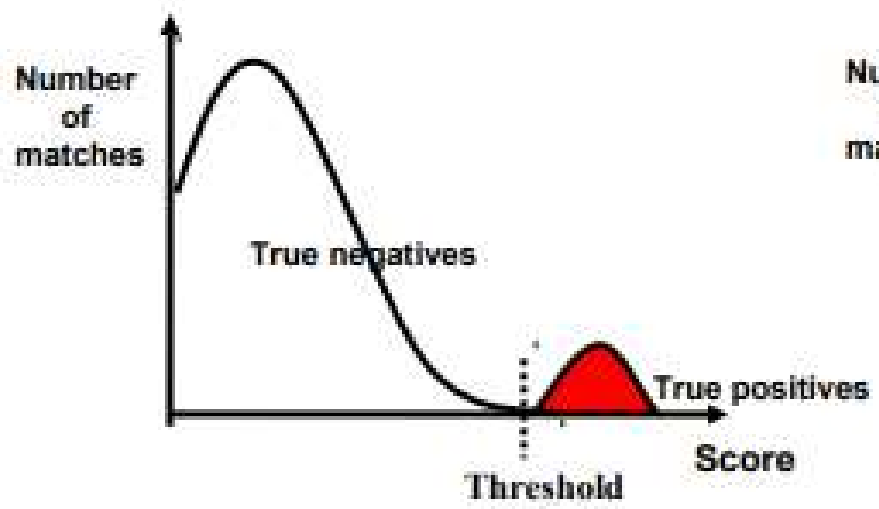


HMMER

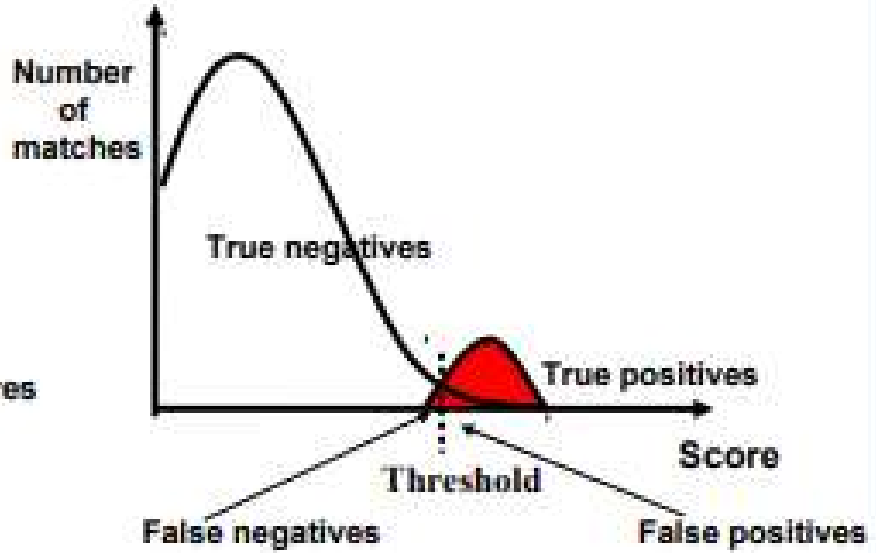
biosequence analysis using profile hidden Markov models

http://www.myoops.org/twocw/mit/NR/rdonlyres/Electrical-Engineering

There is always a chance that a prediction of a feature by a tool is false



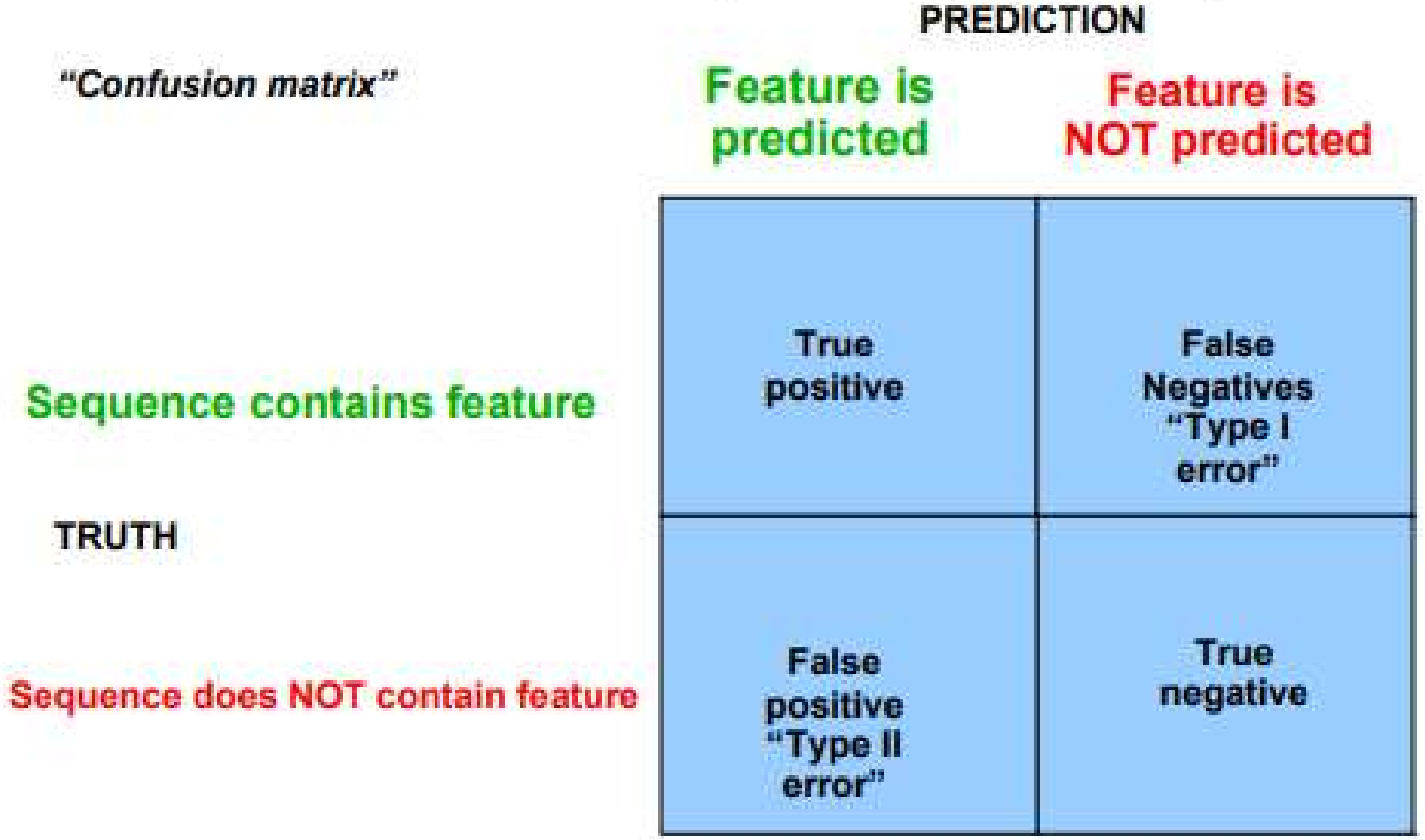
Ideal situation



Reality of the databases

Assessing the performance of categorizing tools with sensitivity and specificity

"Confusion matrix"



Now that we have had a brief overview of how to use the different alignment tools to predict homologs and annotate our genomes,

Let’s Practice!!!!

https://github.com/GlobalInvertebrateGenomicsAlliance/GIGAIII_bioinformatics_workshop/blob/master/Lessons_Day_1/Introduction_to_Annotation_Rhodes.md