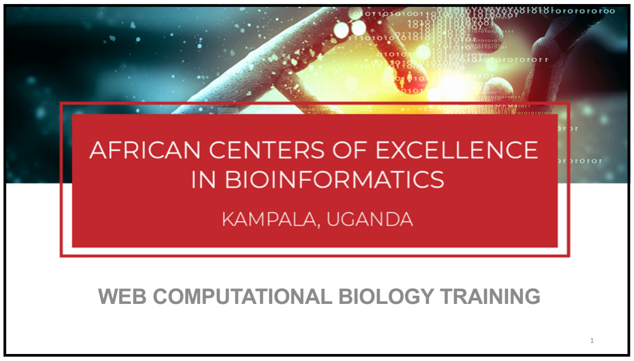
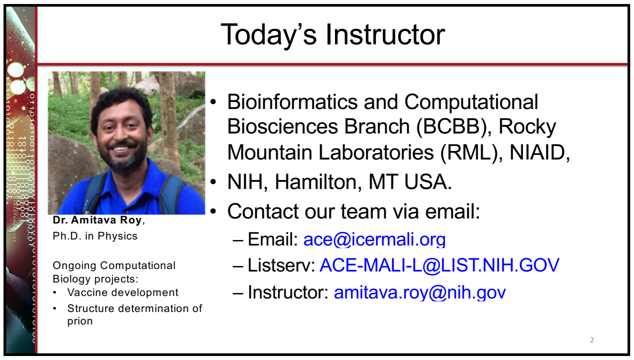


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

WEB COMPUTATIONAL BIOLOGY TRAINING

1



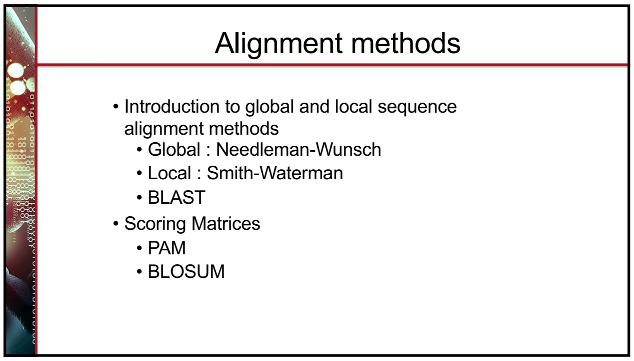
Today's Instructor

Dr. Amitava Roy,
Ph.D. in Physics

Ongoing Computational Biology projects:
- Vaccine development
- Structure determination of prion

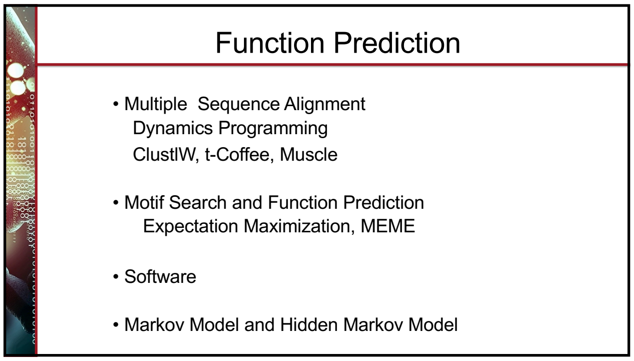
- Bioinformatics and Computational Biosciences Branch (BCBB), Rocky Mountain Laboratories (RML), NIAID,
- NIH, Hamilton, MT USA.
- Contact our team via email:
  - Email: ace@icermali.org
  - Listserv: ACE-MALI-L@LIST.NIH.GOV
  - Instructor: amitava.roy@nih.gov

2



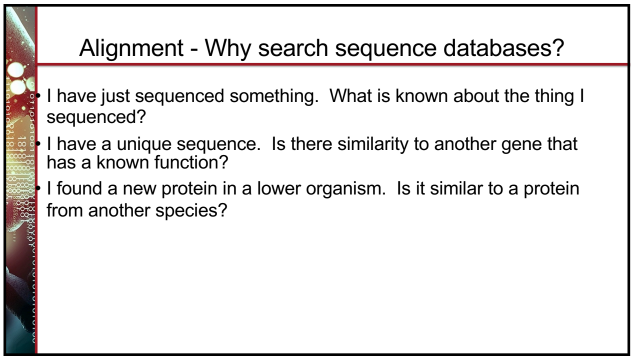
Alignment methods

- Introduction to global and local sequence alignment methods
  - Global : Needleman-Wunsch
  - Local : Smith-Waterman
  - BLAST
- Scoring Matrices
  - PAM
  - BLOSUM



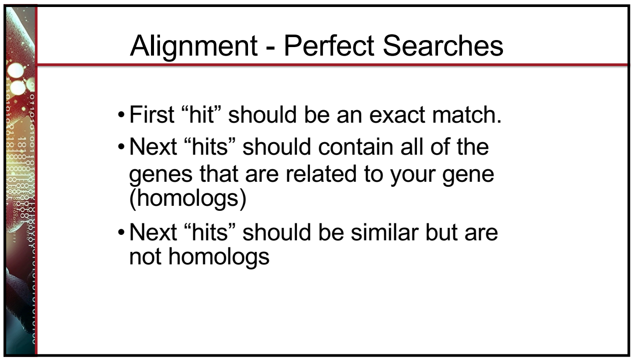
Function Prediction

- Multiple Sequence Alignment
  Dynamics Programming
  ClustlW, t-Coffee, Muscle
- Motif Search and Function Prediction
  Expectation Maximization, MEME
- Software
- Markov Model and Hidden Markov Model



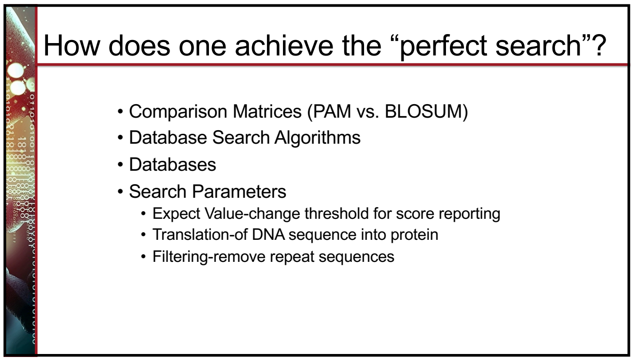
Alignment - Why search sequence databases?

- I have just sequenced something. What is known about the thing I sequenced?
- I have a unique sequence. Is there similarity to another gene that has a known function?
- I found a new protein in a lower organism. Is it similar to a protein from another species?



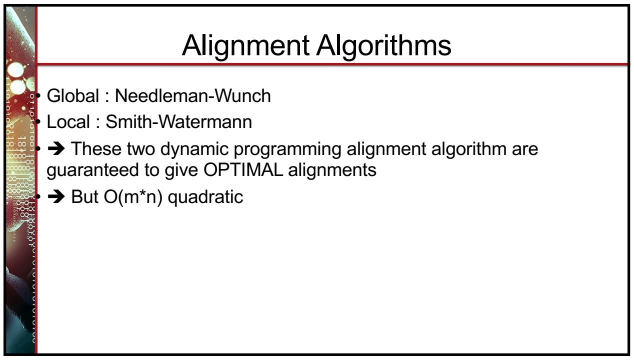
Alignment - Perfect Searches

- First “hit” should be an exact match.
- Next “hits” should contain all of the genes that are related to your gene (homologs)
- Next “hits” should be similar but are not homologs



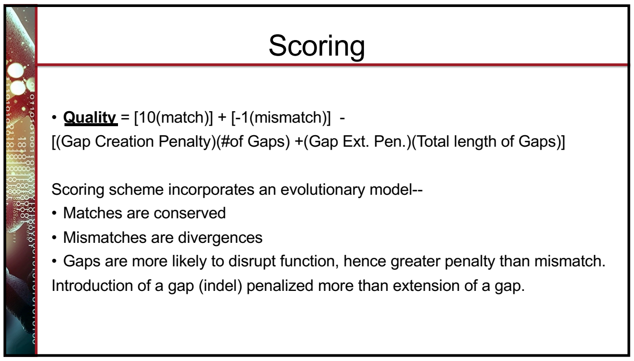
How does one achieve the “perfect search”?

- Comparison Matrices (PAM vs. BLOSUM)
- Database Search Algorithms
- Databases
- Search Parameters
  - Expect Value-change threshold for score reporting
  - Translation-of DNA sequence into protein
  - Filtering-remove repeat sequences



Alignment Algorithms

- Global : Needleman-Wunch
- Local : Smith-Watermann
- ➔ These two dynamic programming alignment algorithm are guaranteed to give OPTIMAL alignments
- ➔ But O(m*n) quadratic

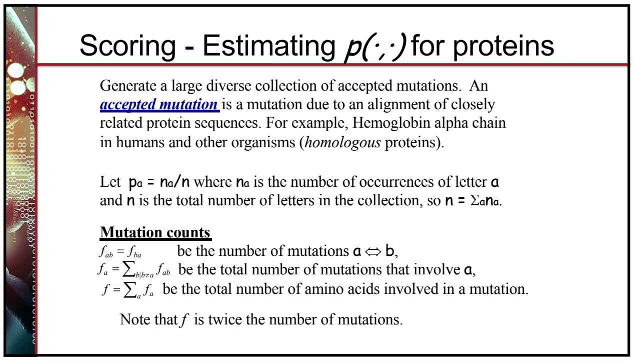


Scoring

- **Quality** = [10(match)] + [-1(mismatch)] - [(Gap Creation Penalty)(#of Gaps) +(Gap Ext. Pen.)(Total length of Gaps)]

Scoring scheme incorporates an evolutionary model--

- Matches are conserved
- Mismatches are divergences
- Gaps are more likely to disrupt function, hence greater penalty than mismatch. Introduction of a gap (indel) penalized more than extension of a gap.



Scoring - Estimating $p(\cdot,\cdot)$ for proteins

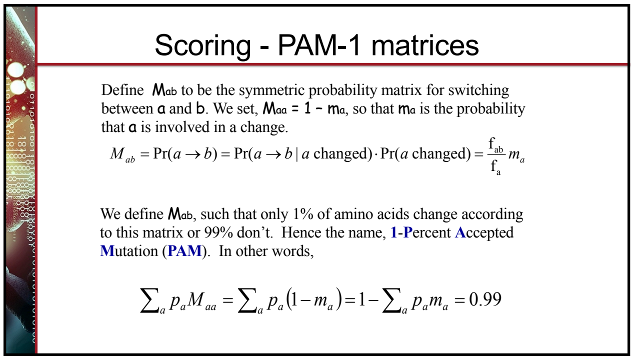
Generate a large diverse collection of accepted mutations. An ***accepted mutation*** is a mutation due to an alignment of closely related protein sequences. For example, Hemoglobin alpha chain in humans and other organisms (*homologous* proteins).

Let $p_a = n_a/n$ where n_a is the number of occurrences of letter a and n is the total number of letters in the collection, so $n = \Sigma_a n_a$.

Mutation counts

$f_{ab} = f_{ba}$ be the number of mutations a ⇔ b,
$f_a = \sum_{b|b \neq a} f_{ab}$ be the total number of mutations that involve a,
$f = \sum_a f_a$ be the total number of amino acids involved in a mutation.

Note that f is twice the number of mutations.



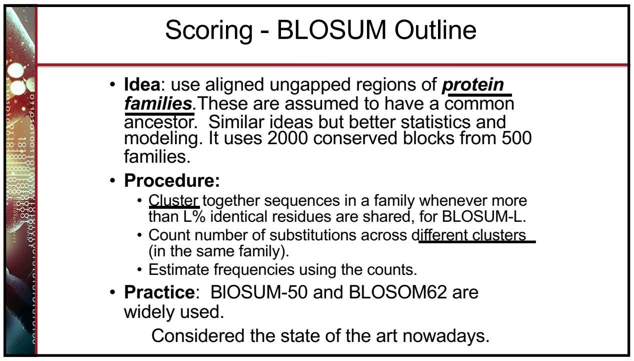
Scoring - PAM-1 matrices

Define M_{ab} to be the symmetric probability matrix for switching between a and b. We set, $M_{aa} = 1 - m_a$, so that m_a is the probability that a is involved in a change.

$$M_{ab} = \Pr(a \to b) = \Pr(a \to b \mid a \text{ changed}) \cdot \Pr(a \text{ changed}) = \frac{\mathrm{f}_{\mathrm{ab}}}{\mathrm{f}_{\mathrm{a}}} m_a$$

We define M_{ab}, such that only 1% of amino acids change according to this matrix or 99% don't. Hence the name, **1-P**ercent **A**ccepted **M**utation (**PAM**). In other words,

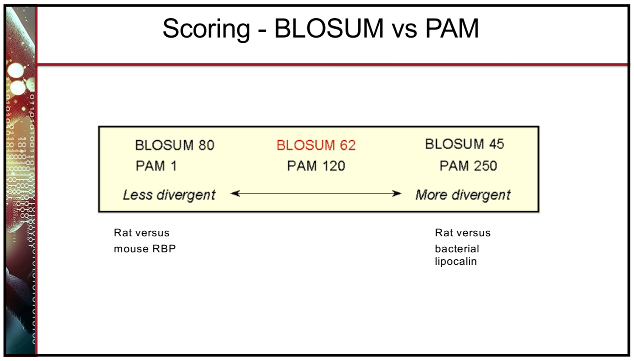
$$\sum_a p_a M_{aa} = \sum_a p_a (1 - m_a) = 1 - \sum_a p_a m_a = 0.99$$



Scoring - BLOSUM Outline

- **Idea**: use aligned ungapped regions of ***protein families***. These are assumed to have a common ancestor. Similar ideas but better statistics and modeling. It uses 2000 conserved blocks from 500 families.
- **Procedure:**
  - Cluster together sequences in a family whenever more than L% identical residues are shared, for BLOSUM-L.
  - Count number of substitutions across different clusters (in the same family).
  - Estimate frequencies using the counts.
- **Practice**: BlOSUM-50 and BLOSOM62 are widely used.

Considered the state of the art nowadays.

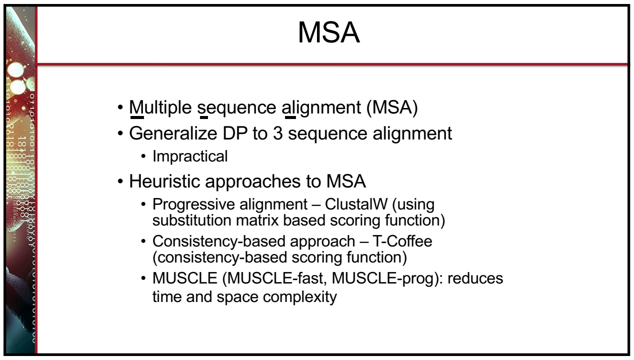


Scoring - BLOSUM vs PAM

BLOSUM 80 PAM 1	BLOSUM 62 PAM 120	BLOSUM 45 PAM 250
Less divergent ←		→ *More divergent*

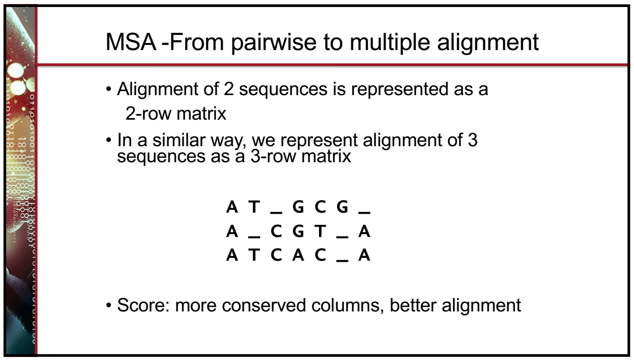
Rat versus mouse RBP

Rat versus bacterial lipocalin



MSA

- Multiple sequence alignment (MSA)
- Generalize DP to 3 sequence alignment
  - Impractical
- Heuristic approaches to MSA
  - Progressive alignment – ClustalW (using substitution matrix based scoring function)
  - Consistency-based approach – T-Coffee (consistency-based scoring function)
  - MUSCLE (MUSCLE-fast, MUSCLE-prog): reduces time and space complexity



MSA -From pairwise to multiple alignment

- Alignment of 2 sequences is represented as a 2-row matrix
- In a similar way, we represent alignment of 3 sequences as a 3-row matrix

```
A T _ G C G _
A _ C G T _ A
A T C A C _ A
```

- Score: more conserved columns, better alignment

What's multiple sequence alignment (MSA)

- A model
- Indicates relationship between residues of different sequences
- Reveals similarity/disimilarity

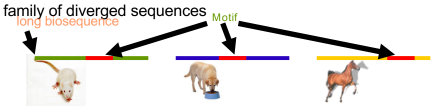
Why we need MSA

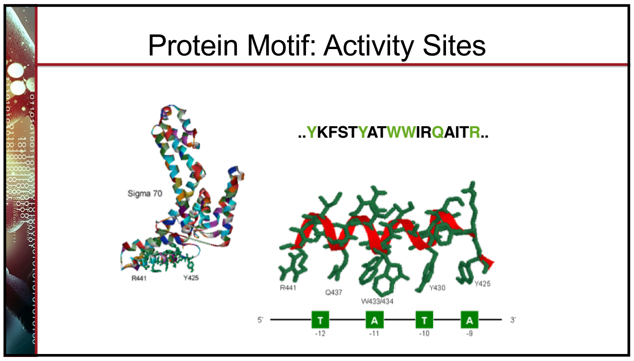
- MSA is central to many bioinformatics applications
- Phylogenetic tree
- Motifs
- Patterns
- Structure prediction (RNA, protein)

Motif

- Set of similar substrings,
  - within a single long sequence

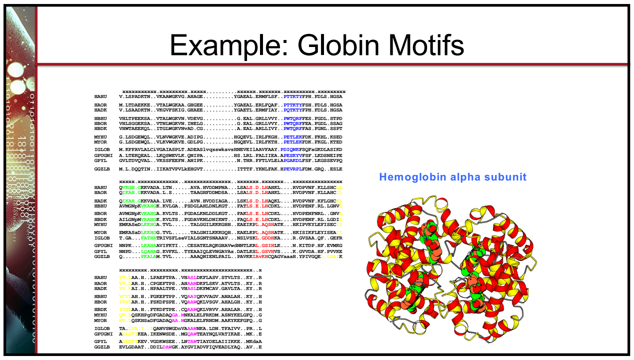
  - or a family of diverged sequences





Protein Motif: Activity Sites

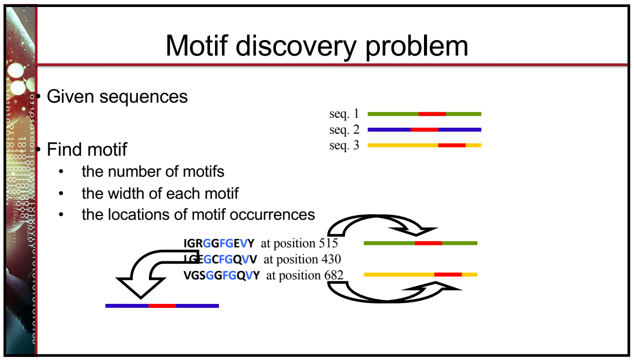
..YKFSTYATWWIRQAITR..



Example: Globin Motifs

```
        xxxxxxxxxxx.xxxxxxxxx.xxxxx.........xxxxxx.xxxxxxx.xxxxxxxxxx.xxxxxxxxx
HAHU    V.LSPADKTN..VKAAWGKVG.AHAGE.........YGAEAL.ERMFLSF..PTTKTYFPH.FDLS.HGSA

HAOR    M.LTDAEKKE..VTALWGKAA.GHGEE.........YGAEAL.ERLFQAF..PTTKTYFSH.FDLS.HGSA
HADK    V.LSAADKTN..VKGVFSKIG.GHAEE.........YGAETL.ERMFIAY..PQTKTYFPH.FDLS.HGSA

HBHU    VHLTPEEKSA..VTALWGKVN.VDEVG..........G.EAL.GRLLVVY..PWTQRFFES.FGDL.STPD
HBOR    VHLSGGEKSA..VTNLWGKVN.INELG..........G.EAL.GRLLVVY..PWTQRFFEA.FGDL.SSAG
HBDK    VHWTAEEKQL..ITGLWGKVNvAD.CG..........A.EAL.ARLLIVY..PWTQRFFAS.FGNL.SSPT

MYHU    G.LSDGEWQL..VLNVWGKVE.ADIPG.........HGQEVL.IRLFKGH..PETLEKFDK.FKHL.KSED
MYOR    G.LSDGEWQL..VLKVWGKVE.GDLPG.........HGQEVL.IRLFKTH..PETLEKFDK.FKGL.KTED

IGLOB   M.KFFAVLALCiVGAIASPLT.ADEASlvqsswkavsHNEVEIlAAVFAAY.PDIQNKFSQFaGKDLASIKD
GPUGNI  A.LTEKQEAL..LKQSWEVLK.QNIPA.........HS.LRL.FALIIEA.APESKYVFSF.LKDSNEIPE
GPYL    GVLTDVQVAL..VKSSFEEFN.ANIPK..........N.THR.FFTLVLEiAPGAKDLFSF.LKGSSEVPQ

GGZLB   M.L.DQQTIN..IIKATVPVLkEHGVT..........ITTTF.YKNLFAK.HPEVRPLFDM.GRQ..ESLE


        xxxxx.xxxxxxxxxxxx..xxxxxxxxxxxxxx..xxxxxxx.xxxxxxx...xxxxxxxxxxxxxxx
HAHU    QVKGH.GKKVADA.LTN......AVA.HVDDMPNA...LSALS.D.LHAHKL....RVDPVNF.KLLSHCLL
HAOR    QIKAH.GKKVADA.L.S......TAAGHFDDMDSA...LSALS.D.LHAHKL....RVDPVNF.KLLAHCIL

HADK    QIKAH.GKKVAAA.LVE......AVN.HVDDIAGA...LSKLS.D.LHAQKL....RVDPVNF.KFLGHCFL
HBHU    AVMGNpKVKAHGK.KVLGA..FSDGLAHLDNLKGT...FATLS.E.LHCDKL....HVDPENF.RL.LGNVL

HBOR    AVMGNpKVKAHGA.KVLTS..FGDALKNLDDLKGT...FAKLS.E.LHCDKL....HVDPENFNRL..GNVL
HBDK    AILGNpMVRAHGK.KVLTS..FGDAVKNLDNIKNT...FAQLS.E.LHCDKL....HVDPENF.RL.LGDIL
MYHU    EMKASeDLKKHGA.TVL......TALGGILKKKGHH..EAEIKPL.AQSHATK...HKIPVKYLEFISECII

MYOR    EMKASaDLKKHGG.TVL......TALGNILKKKGQH..EAELKPL.AQSHATK...HKISIKFLEYISEAII
IGLOB   T.GA...FATHATRIVSFLseVIALSGNTSNAAAV...NSLVSKL.GDDHKA....R.GVSAA.QF..GEFR

GPUGNI  NNPK...LKAHAAVIFKTI...CESATELRQKGHAVwdNNTLKRL.GSIHLK....N.KITDP.HF.EVMKG
GPYL    NNPD...LQAHAG.KVFKL..TYEAAIQLEVNGAVAs.DATLKSL.GSVHVS....K.GVVDA.HF.PVVKE
GGZLB   Q......PKALAM.TVL......AAAQNIENLPAIL..PAVKKIAvKHCQAGVaaaH.YPIVGQEL.LGAIK


        xxxxxxxxx.xxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxx..x
HAHU    VT.LAA.H..LPAEFTPA..VHASLDKFLASV.STVLTS..KY..R
HAOR    VV.LAR.H..CPGEFTPS..AHAAMDKFLSKV.ATVLTS..KY..R
HADK    VV.VAI.H..HPAALTPE..VHASLDKFMCAV.GAVLTA..KY..R

HBHU    VCVLAH.H..FGKEFTPP..VQAAYQKVVAGV.ANALAH..KY..H
HBOR    IVVLAR.H..FSKDFSPE..VQAAWQKLVSGV.AHALGH..KY..H

HBDK    IIVLAA.H..FTKDFTPE..CQAAWQKLVRVV.AHALAR..KY..H
MYHU    QV.LQSKHPgDFGADAQGA.MNKALELFRKDM.ASNYKELGFQ..G
MYOR    HV.LQSKHSaDFGADAQAA.MGKALELFRNDM.AAKYKEFGFQ..G

IGLOB   TA.LVA.Y..LQANVSWGDnVAAAWNKA.LDN.TFAIVV..PR..L
GPUGNI  ALLGTIKEA.IKENWSDE..MGQAWTEAYNQLVATIKAE..MK..E

GPYL    AILKTIKEV.VGDKWSEE..LNTAWTIAYDELAIIIKKE..MKdaA
GGZLB   EVLGDAAT..DDILDAWGK.AYGVIADVFIQVEADLYAQ..AV..E
```

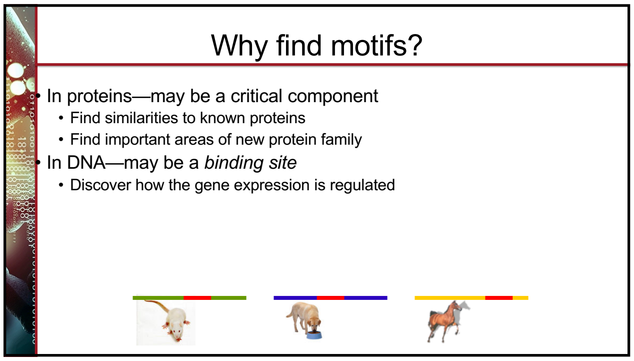
Hemoglobin alpha subunit



Motif discovery problem

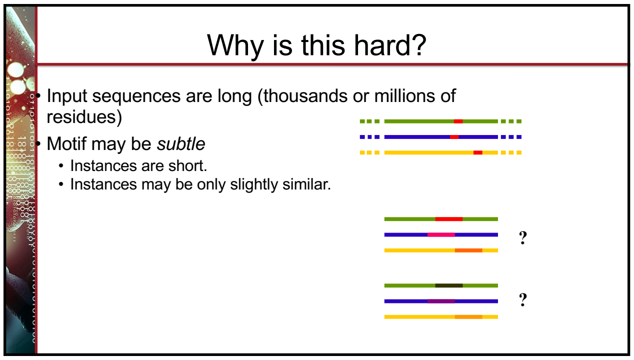
- Given sequences
- Find motif
  - the number of motifs
  - the width of each motif
  - the locations of motif occurrences

IGRGGFGEVY at position 515
LGEGCFGQVV at position 430
VGSGGFGQVY at position 682



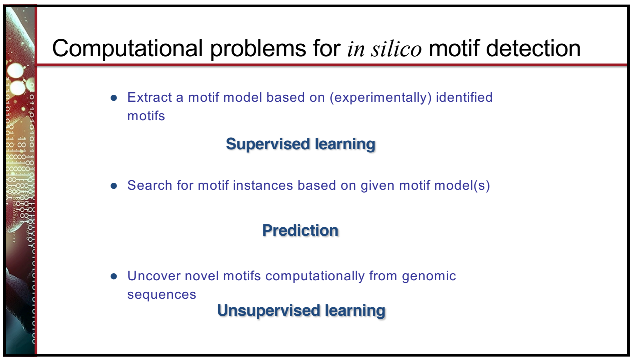
Why find motifs?

- In proteins—may be a critical component
  - Find similarities to known proteins
  - Find important areas of new protein family
- In DNA—may be a *binding site*
  - Discover how the gene expression is regulated



Why is this hard?

- Input sequences are long (thousands or millions of residues)
- Motif may be *subtle*
  - Instances are short.
  - Instances may be only slightly similar.



Computational problems for *in silico* motif detection

- Extract a motif model based on (experimentally) identified motifs

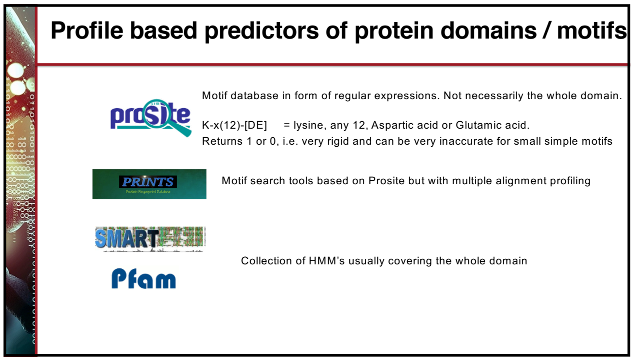
Supervised learning

- Search for motif instances based on given motif model(s)

Prediction

- Uncover novel motifs computationally from genomic sequences

Unsupervised learning



Profile based predictors of protein domains / motifs

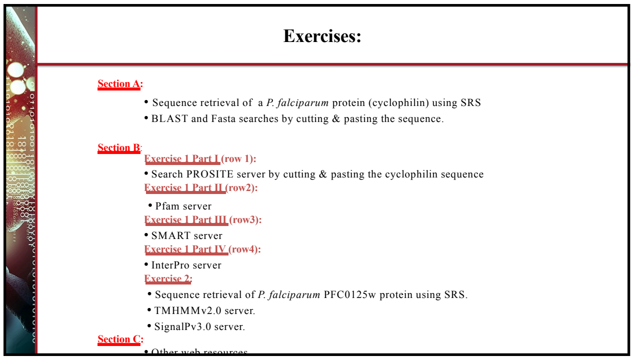
Motif database in form of regular expressions. Not necessarily the whole domain.

K-x(12)-[DE] = lysine, any 12, Aspartic acid or Glutamic acid.
Returns 1 or 0, i.e. very rigid and can be very inaccurate for small simple motifs

Motif search tools based on Prosite but with multiple alignment profiling

Collection of HMM’s usually covering the whole domain

prosite
PRINTS
Protein Fingerprint Database
SMART
Pfam



Exercises:

Section A:

- Sequence retrieval of a *P. falciparum* protein (cyclophilin) using SRS
- BLAST and Fasta searches by cutting & pasting the sequence.

Section B:

Exercise 1 Part I (row 1):

- Search PROSITE server by cutting & pasting the cyclophilin sequence

Exercise 1 Part II (row2):

- Pfam server

Exercise 1 Part III (row3):

- SMART server

Exercise 1 Part IV (row4):

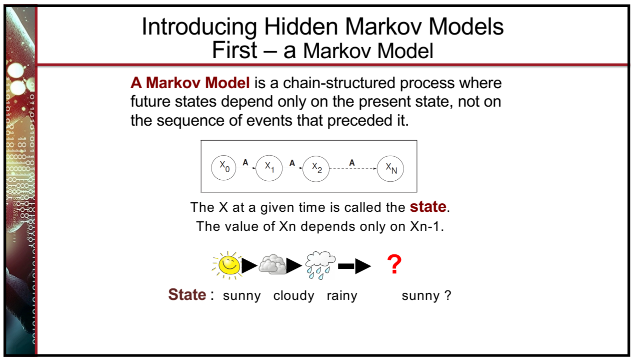
- InterPro server

Exercise 2:

- Sequence retrieval of *P. falciparum* PFC0125w protein using SRS.
- TMHMMv2.0 server.
- SignalPv3.0 server.

Section C:

- Other web resources

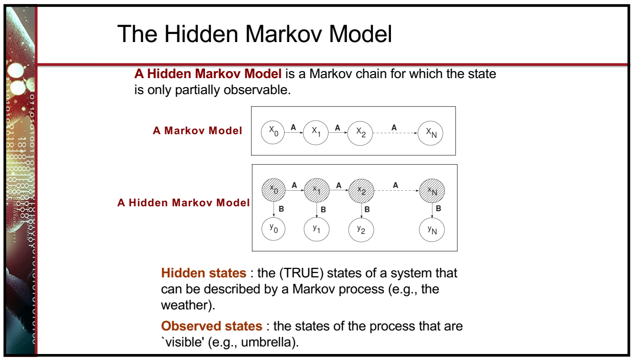


Introducing Hidden Markov Models
First – a Markov Model

A Markov Model is a chain-structured process where future states depend only on the present state, not on the sequence of events that preceded it.

The X at a given time is called the **state**.
The value of Xn depends only on Xn-1.

State : sunny cloudy rainy sunny ?



The Hidden Markov Model

A Hidden Markov Model is a Markov chain for which the state is only partially observable.

A Markov Model

A Hidden Markov Model

Hidden states : the (TRUE) states of a system that can be described by a Markov process (e.g., the weather).

Observed states : the states of the process that are `visible' (e.g., umbrella).