SESSION 10. A FUNCTION TO EVERY GENE

Royal blood and order in the sequence universe

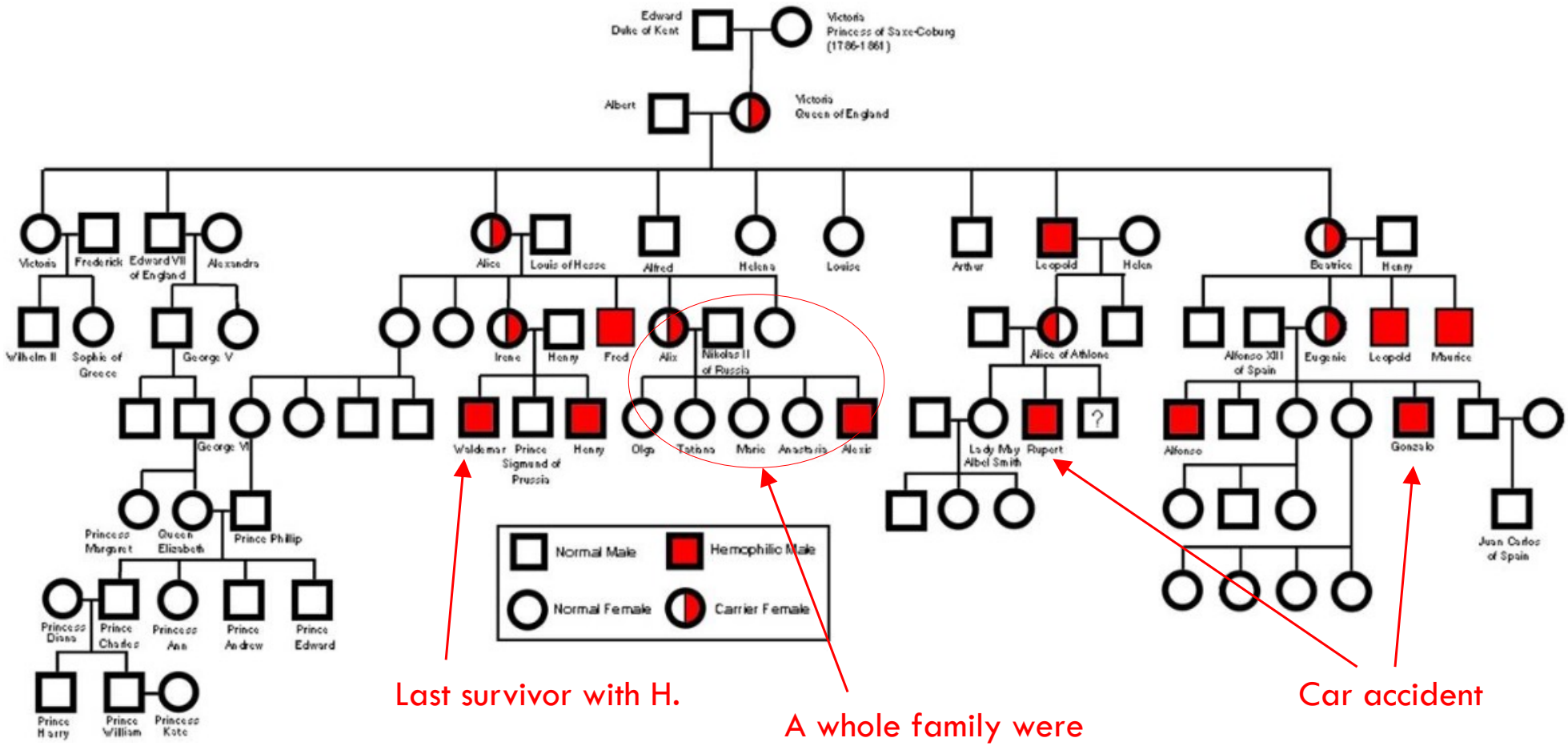
A letter written by Yevgeny Botkin, the family doctor of Russian royal family

The boy is in such indescribable pain day and night that no one from among his closest relatives, though they do not spare themselves, has the strength to bear looking after him too long, not to mention his mother, with her chronically ill heart... ....

Haemophilia in Russian royal family

- *Alexei, the son of the Nicholas II and Alexandra and the successor to the throne of all Russians, suffered from* ***a congenital bleeding disease (Haemophilia)***
- *The defect is a* ***missing or malfunctioning blood-clotting factor.***
- *The particular Haemophilia that affected Alexei is a genetic disease* ***linked to the X.***
- *For this reason, it is more prevalent in males.*

Haemophilia in England, Spain, and Russian royal families

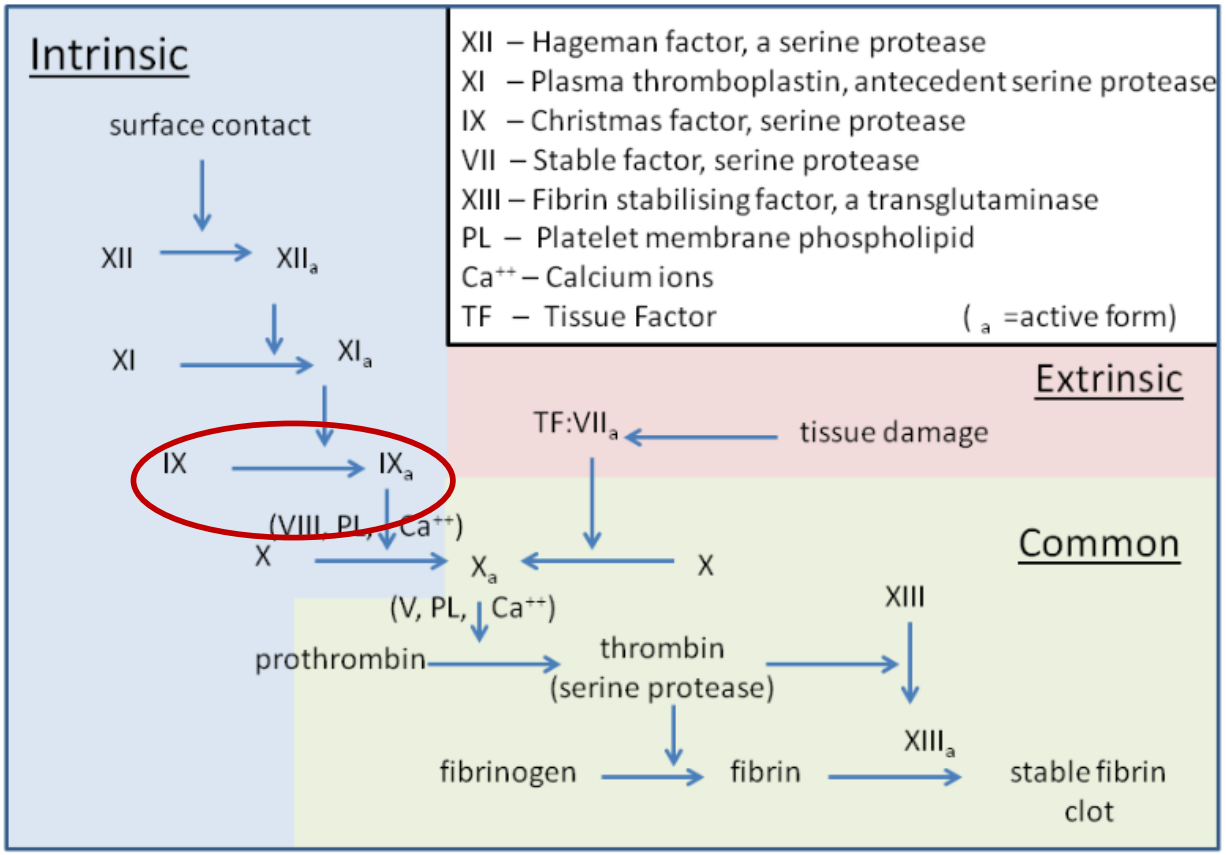


More details about Haemophilia

- **Haemophilia A –** Deficiency in the activity of coagulation factor VIII.
- **Haemophilia B** (Christmas disease) deficiency in the activity of factor IX.
- Both A and B are X-linked
- Based on the pedigree, Alexei was likely to have had either one.
- In 2009 study, found that **Alexei suffered from haemophilia B.**
- The mutation in factor IX alters RNA splicing which leads to production of truncated form.

Blood-clotting factors

The three pathways that makeup the classical blood coagulation pathway



Enzymatic cascade → at each step a molecular signal is amplified → Efficient and rapid response to an early trauma.

Each enzyme has proteolytic activity

Extrinsic and Intrinsic pathways are related each other in their **biochemical properties and their sequence and domain structure.**

→ They are related by evolution (gene duplication)

Genotype Analysis Identifies the Cause of the "Royal Disease"

Evgeny I. Rogaev,[1,2,3,4]*† Anastasia P. Grigorenko,[1,2,3]* Gulnaz Faskhutdinova,[1] Ellen L. W. Kittler,[1] Yuri K. Moliaka[1]

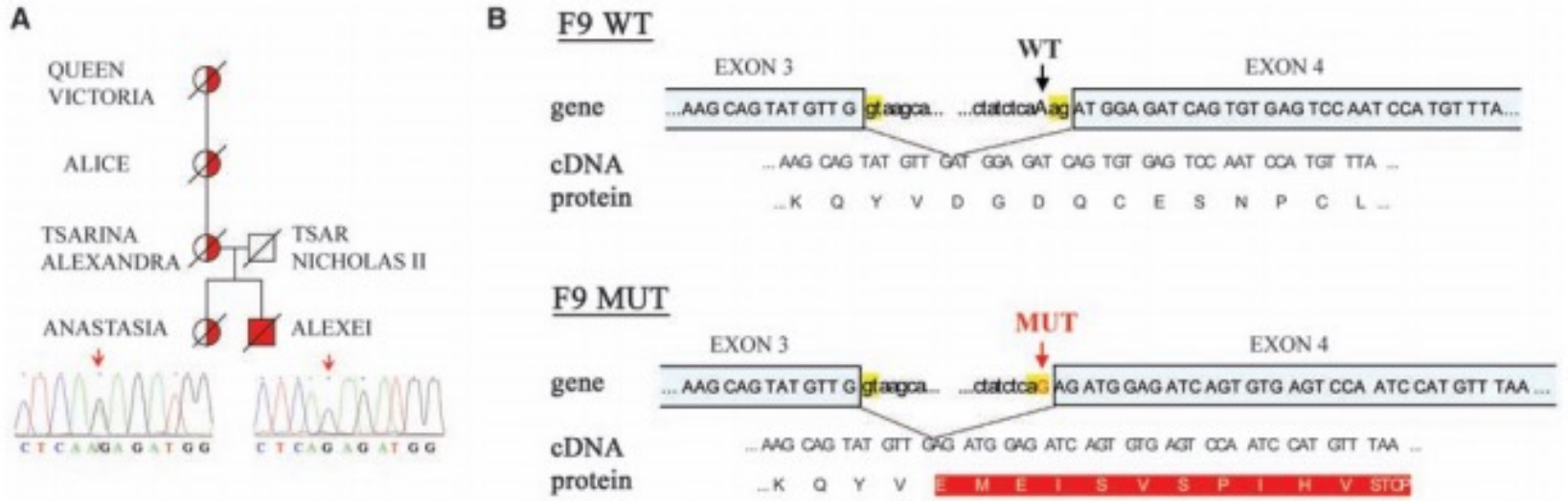
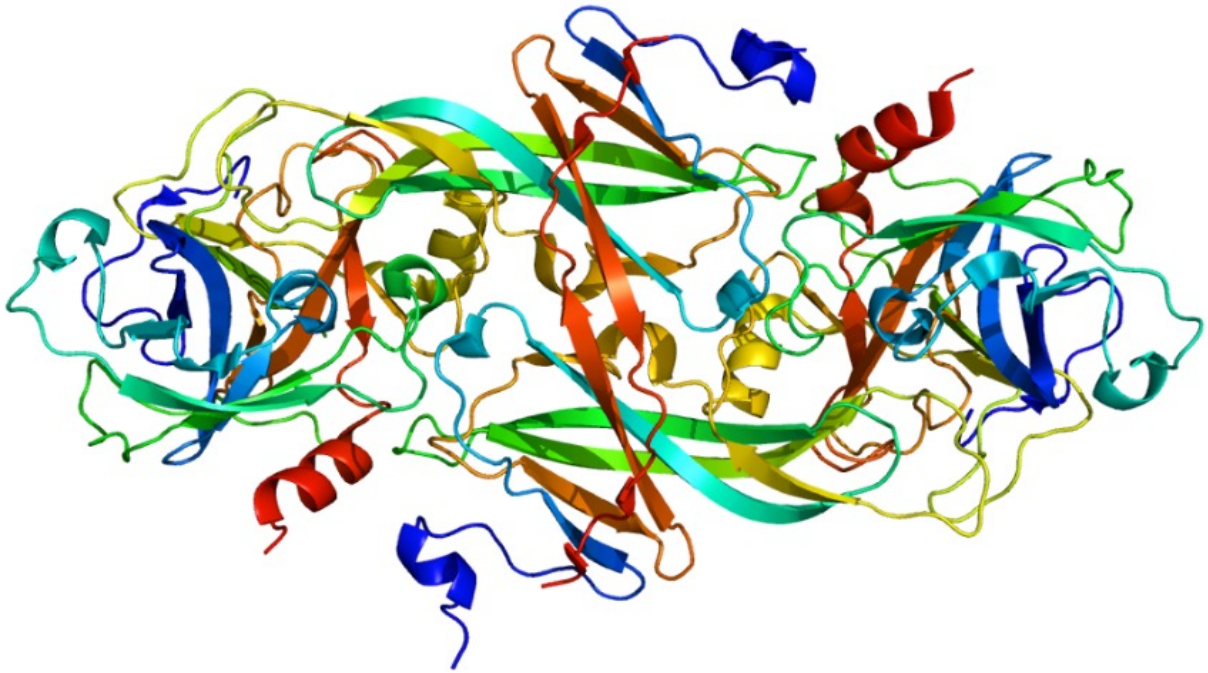


Fig. 1. The royal disease was likely caused by a point mutation in *F9*, a gene on the X chromosome that encodes blood coagulation factor IX. (**A**) Partial pedigree of the royal family, showing transmission of the mutation from Queen Victoria to Empress Alexandra and from Alexandra to Prince Alexei, her hemophilic son. Alexei was hemizygous for the mutation, whereas Alexandra and one of her daughters, putative Grand Duchess Anastasia, were heterozygous carriers. (**B**) The A-to-G mutation occurs just upstream of exon 4 in the *F9* gene and is predicted to create a new splice acceptor site that could lead to production of a truncated factor IX protein (*6*). WT indicates wild type.

Protein domain architecture of blood clotting enzymes

Structure of blood coagulation factor IX



- ***Four PAN domains***
  → *Mediate a number of interaction with other proteins*
- ***One serine-protease domain***
  → *Proteolysis of factor X*

http://pfam.xfam.org

http://www.rcsb.org/pdb/home/home.do

Protein domain architecture

- A majority of proteins are built from more than one domain.
- Domain architecture – a sequence of protein domains.
- In the view of evolution, the domains are like Lego pieces to assemble something.

- Chromosomal rearrangement promotes the evolution of proteins in the domain levels → Chimeric genes…

Bioinformatics of protein domains

BLAST is not enough sensitive to identify homologous domains because the aa sequences can be rapidly evolved as long as they maintain their 3D structures.

Therefore, protein sequences may have as little as 5-10% of sequence identity, although they are evolutionarily related.

- → A suggested approach: based on profiles (or position-specific scoring matrices, PSSMs)
- → For a family of evolutionarily related protein sequences., a statistical model (or PSSM) from MSA is created.

PSSM (or PWM)

A matrix of vectors of the size *k* (4 for NT; 20 for aa) X the sequence length.

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

the corresponding PFM is:

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}.$$

and therefore the resulting PPM is:

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

$b_k = 1/|k|$ (k=4 for NT; k=20 for aa)

or

$$b_k = (\sum_{j=1}^{n} M_{k,j})/n$$

$$M_{k,j} = \ln\left(M_{k,j}/b_k\right).$$

PWM

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.18 & 0.87 & -0.91 & -\infty & -\infty & 0.87 & 1.02 & -0.22 & -0.91 \\ -0.22 & -0.22 & -0.91 & -\infty & -\infty & -0.22 & -0.91 & -0.91 & -0.22 \\ -0.91 & -0.91 & 1.02 & 1.38 & -\infty & -0.91 & -0.91 & 0.69 & -0.91 \\ 0.47 & -0.91 & -0.91 & -\infty & 1.38 & -0.91 & -0.91 & -0.22 & 0.87 \end{bmatrix}.$$

X='AAGGTTGGC'

S(X|M) = 0.18+0.87+1.02+1.38+1.38-0.91-0.91+0.69-0.22

Use of PSSM (or PWM)

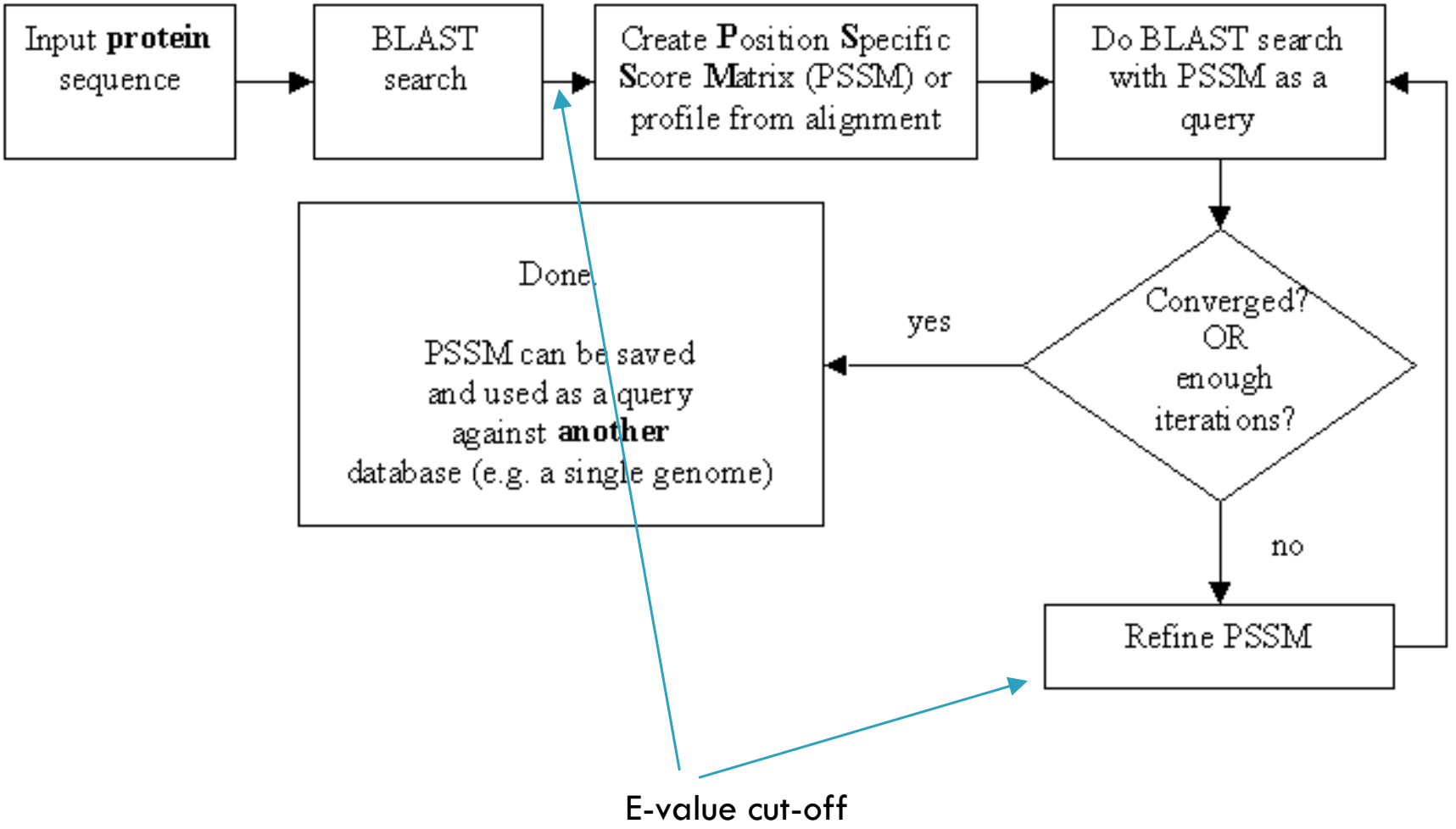
Given an unknown, uncharacterized protein,

We can search it against a library of the PWM profiles, and thereby learn about the structural and functional properties of the protein.

Profile-based search programs

- position-specific iterated BLAST (PSI-BLAST; Altschul et al., 1997)
  - Two steps: 1) regular BLAST search against protein DB 2) PSI-BLAST using a profile from hits and iteration of 2)
- Profile Hidden Markov Models (HMMER; Eddy et al., 1998)
  - A formal probabilistic framework of PSI-like search

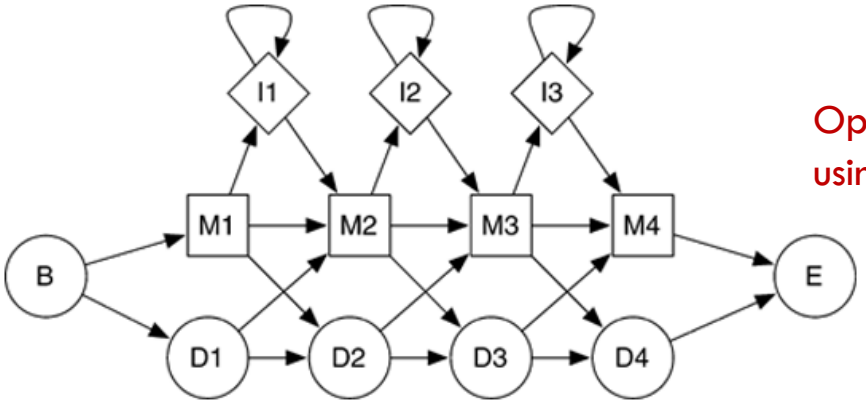
PSI-BLAST


Input **protein** sequence
BLAST search
Create **P**osition **S**pecific **S**core **M**atrix (PSSM) or profile from alignment
Do BLAST search with PSSM as a query
Converged? OR enough iterations?
yes
Done.
PSSM can be saved and used as a query against **another** database (e.g. a single genome)
no
Refine PSSM
E-value cut-off

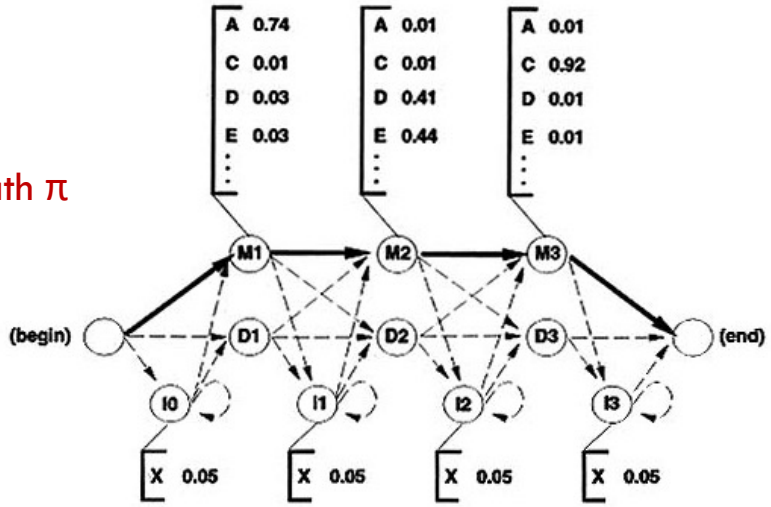
Profiled HMM

```
RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSAIGILS
RYDSRTTIFSPLREGRLYQVEYAMEAISHA.GTCLGILS
RYDSRTTIFSP..EGRLYQVEYAQEAISNA.GTAIGILS
RYDSRTTIFSP..EGRLYQVEYAMEAISHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEAIGHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSALGVLA
RYDSRTTTFSP..EGRLYQVEYALEAINNA.SITIGLIT
SYDSRTTIFSP..EGRLYQVEYALEAINHA.GVALGIVA
```

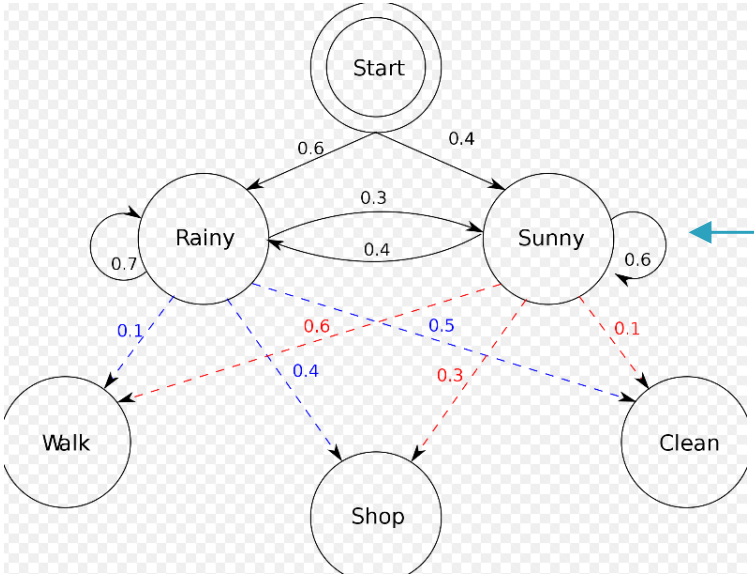
Training HMM



Optimal path π using DP



Hidden Markov Model



Hidden states (latent)

Observable states

WWWCCCCSWW

SSSRRRSSSS

HMMer

Pfam download (HMM profiles)

```
[jwnam@biglab-master Session9]$ wget ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam29.0/Pfam-A.hmm.gz
--2016-05-10 11:28:50--  ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam29.0/Pfam-A.hmm.gz
           => "Pfam-A.hmm.gz"
Resolving ftp.ebi.ac.uk... 193.62.194.182
Connecting to ftp.ebi.ac.uk|193.62.194.182|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/databases/Pfam/releases/Pfam29.0 ... done.
==> SIZE Pfam-A.hmm.gz ... 254622237
==> PASV ... done.    ==> RETR Pfam-A.hmm.gz ... done.
Length: 254622237 (243M) (unauthoritative)

 1% [==>
```

Building HMM profile

```
fastacmd -i bc_seqid.txt -d swissprot >clotting.fa
clustalw2 clotting.fa
hmmbuild clotting.hmm clotting.aln
```

clotting.hmm

```
HMMER3/f [3.1b2 | February 2015]
NAME  clotting
LENG  2076
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Tue May 10 11:51:12 2016
NSEQ  14
EFFN  1.425293
CKSUM 1327816984
STATS LOCAL MSV      -13.6585  0.69432
STATS LOCAL VITERBI  -15.1043  0.69432
STATS LOCAL FORWARD   -7.5991  0.69432
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q        R        S        T        V        W        Y
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.59040  4.11795  2.92923  2.69436  3.26618  2.87839  3.60499  2.90278  2.67671  2.48318  3.65878  3.05735  3.25487  3.06152  2.94023  2.59026  2.81288  2.69550  4.39593  3.40801
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.00000        *
      1   2.96588  4.51530  4.04489  3.63803  3.20586  3.83093  4.37414  2.38103  3.40634  1.80115  1.38600  3.91604  4.29150  3.78848  3.61215  3.31456  3.27110  2.38667  5.04612  3.80753      1 m - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.48576  0.95510
      2   2.87738  4.86301  2.90011  2.65476  4.03102  3.36323  3.81304  3.66112  2.47584  3.16389  4.19185  3.11447  3.91662  1.24152  2.76010  2.92530  3.17434  3.39104  5.30377  3.99723      2 q - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.48576  0.95510
      3   3.02783  4.44770  4.32975  3.94537  3.37134  4.06203  4.67426  1.08311  3.77343  1.95821  3.30732  4.19130  4.49362  4.12214  3.97048  3.55690  3.32706  1.84575  5.26131  3.99926      3 i - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.48576  0.95510
```

HMMer

HMM profile indexing

```
hmmpress clotting.hmm
```

```
[jwnam@biglab-master Session9]$ hmmpress clotting.hmm
Working...    done.
Pressed and indexed 1 HMMs (1 names).
Models pressed into binary file:   clotting.hmm.h3m
SSI index for binary model file:   clotting.hmm.h3i
Profiles (MSV part) pressed into:  clotting.hmm.h3f
Profiles (remainder) pressed into: clotting.hmm.h3p
```

Searching HMM profile with a query sequence

```
hmmscan --domtblout query.tab clotting.hmm query.fa
```

```
#                                                                            --- full sequence ---  -------------- this domain -------------   hmm coord   ali coord   env coord
# target name        accession   tlen query name                    accession   qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score  bias  from    to  from    to  from    to  acc desc
#------------------- ---------- -----          --------------------- ---------- ----- --------- ------ ----- --- --- --------- --------- ------ ----- ----- ----- ----- ----- ----- ----- ---- ----
clotting            -          2076 gi|119766|sp|P08709.1|FA7_HUMAN -            466   3.6e-66  209.4   0.0   1   3     3e-07     3e-07   14.0   0.2   706   820    27   121     8   174 0.74 -
clotting            -          2076 gi|119766|sp|P08709.1|FA7_HUMAN -            466   3.6e-66  209.4   0.0   2   3    0.0013    0.0013    2.0   0.1   879   916   112   149   105   200 0.69 -
clotting            -          2076 gi|119766|sp|P08709.1|FA7_HUMAN -            466   3.6e-66  209.4   0.0   3   3   7.1e-61   7.1e-61  191.9   0.0  1118  1387   204   454   197   459 0.95 -
#
# Program:         hmmscan
# Version:         3.1b2 (February 2015)
# Pipeline mode:   SCAN
# Query file:      query.fa
# Target file:     clotting.hmm
# Option settings: hmmscan --domtblout test.tab clotting.hmm query.fa
# Current dir:     /home/jwnam/CLASS/Computatonal_biology/Session9
# Date:            Tue May 10 11:58:25 2016
# [ok]
```

query.tab

```
# hmmscan :: search sequence(s) against a profile database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query sequence file:             query.fa
# target HMM database:             clotting.hmm
# per-dom hits tabular output:     test.tab
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       gi|119766|sp|P08709.1|FA7_HUMAN  [L=466]
Description: RecName: Full=Coagulation factor VII; AltName: Full=Serum prothrombin conversion accelerator
Flags: Precursor
Scores for complete sequence (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model    Description
    ------- ------ -----    ------- ------ -----   ---- --  -------- -----------
    3.6e-66  209.4   0.0    7.1e-61  191.9   0.0    2.6  3  clotting


Domain annotation for each model (and alignments):
>> clotting
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   14.0   0.2     3e-07     3e-07     706     820 ..      27     121 ..       8     174 .. 0.74
   2 !    2.0   0.1    0.0013    0.0013     879     916 ..     112     149 ..     105     200 .. 0.69
   3 !  191.9   0.0   7.1e-61   7.1e-61    1118    1387 ..     204     454 ..     197     459 .. 0.95
```

```
 == domain 3  score: 191.9 bits;  conditional E-value: 7.1e-61
                          clotting 1118 nqekkiqeeivkketliqetgtklmknllepwivvldtstqnmkhlcggtLiqedyvlkaahcltqsplsdsltvv 1193
                                        ++ +k+q +iv+++ +   +g++       pw+v+l ++     +lcggtLi++ +v++aahc+++ +++++l +v
 gi|119766|sp|P08709.1|FA7_HUMAN  204 RNASKPQGRIVGGKVCP--KGEC-------PWQVLLLVN---GAQLCGGTLINTIWVVSAAHCFDKIKNWRNLIAV 267
                                        56678999999999998..6777.......*******95...58******************************** PP

                          clotting 1194 lgqdnsseleaasqkkkveevqesseflqgakknnldlalLkLemtgdrevvssdtvatislpskkventvlpkpt 1269
                                        lg+++ se++ ++q+++v +v++ s++  g+  +n+d+alL+L+    ++vv++d+v +++lp++++++l+  +
 gi|119766|sp|P08709.1|FA7_HUMAN  268 LGEHDLSEHDGDEQSRRVAQVIIPSTYVPGT--TNHDIALLRLH----QPVVLTDHVVPLCLPERTFSERTLAFVR 337
                                        *******************************9..9**********....*************************** PP

                          clotting 1270 kckvsGwgellekveiyqpghLklvevsllqneecakke...essakitskllcagakdnhipkdeckgdsggplk 1342
                                         + vsGwg+ll+++ +  + +L +++v+ l +++c +++    +s +it+ ++cag++d+   kd+ckgdsggp++
 gi|119766|sp|P08709.1|FA7_HUMAN  338 FSLVSGWGQLLDRGAT--ALELMVLNVPRLMTQDCLQQSrkvGDSPNITEYMFCAGYSDGS--KDSCKGDSGGPHA 409
                                        **************99..69***************99996667899*************77..************* PP

                          clotting 1343 takkdtillhgivsigegcakkekegvytkvgrtekwisqnpkvl 1387
                                        t+ ++t++l+givs+g+gca+ ++ gvyt+v+++ +w+++ ++++
 gi|119766|sp|P08709.1|FA7_HUMAN  410 THYRGTWYLTGIVSWGQGCATVGHFGVYTRVSQYIEWLQKLMRSE 454
                                        ******************************************99986 PP
```

parse_hmmscan.py

```
#!/usr/bin/python

import re
import sys

filename = sys.argv[1]

print 'protname\tlen\tdomname\tbegin\tend'

for line in open(filename):
    if not re.search('^\#', line):  # avoid all lines beginning
                                    # with the '#' character

        col = re.split(' +', line)

        domname =  col[0]
        protname = col[3]

        protname = re.sub('.*\|', '', protname)
        length   = col[5]
        evalue   = float(col[12])
        begin    = col[17]
        end      = col[18]
        if evalue < 1e-5:
            print protname, '\t', length, '\t', domname,
            print '\t', begin, '\t', end
```

```
[jwnam@biglab-master Session9]$ python parse_hmmscan.py test.tab
protname      len           domname       begin         end
FA7_HUMAN     466           clotting      27            121
FA7_HUMAN     466           clotting      204           454
```