

# SESSION 10. A FUNCTION TO EVERY GENE

**Royal blood and order in  
the sequence universe**



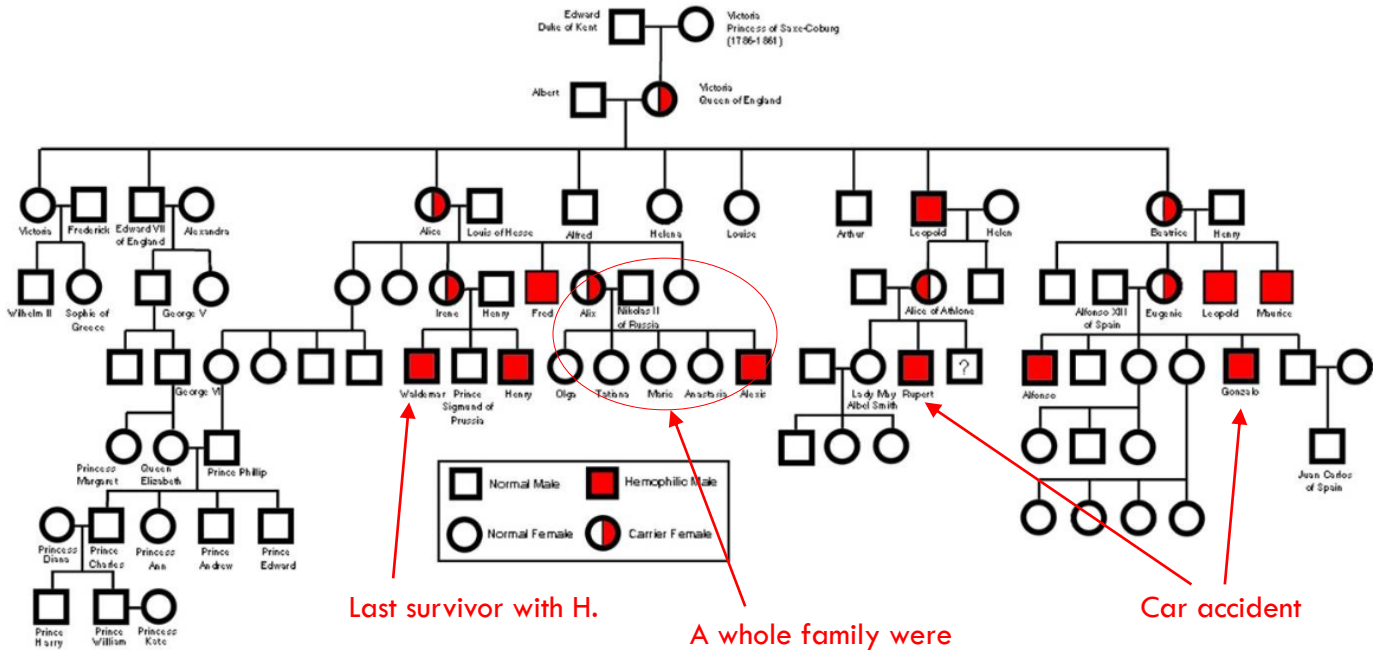
## **A letter written by Yevgeny Botkin, the family doctor of Russian royal family**

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

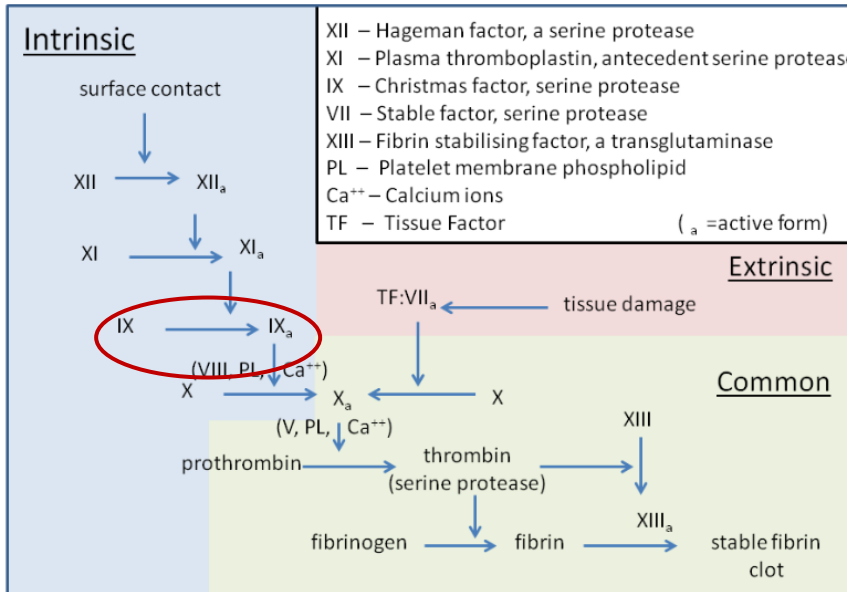


## More details about Haemophilia

- **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)



# 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.

the corresponding PFM is:

$$b_k = 1/|k| \quad (k=4 \text{ for NT}; k=20 \text{ for aa})$$

GAGGTAAC  
TCCGTAAGT  
CAGGTTGGA  
ACAGTCAGT  
TAGGTCATT  
TAGGTAAGT  
ATGGTAACT  
CAGGTATAC  
TGTGTGAGT  
AAGGTAAGT



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

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

or

and therefore the resulting PPM is:

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

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

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

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)

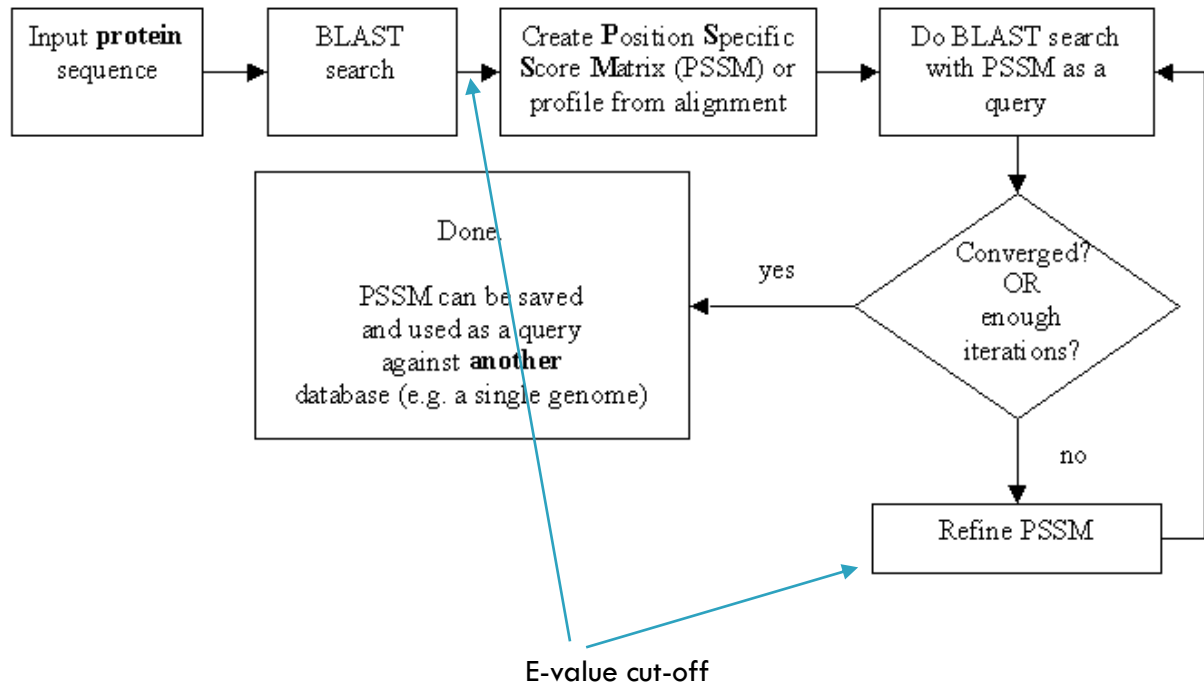
*Given an unknown, uncharacterized protein,*

*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

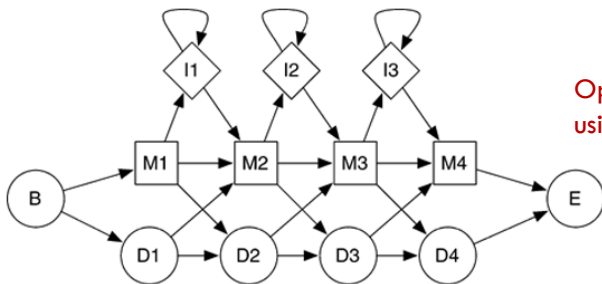


# Profiled HMM

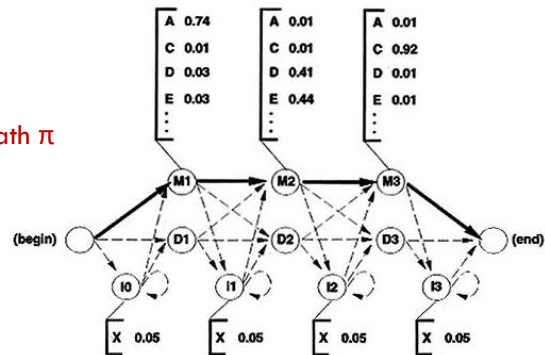
```

RYDSRTTIFSP..EGRLYQVEYAMEEAIGNA.GSAIGTILS
RYDSRTTIFSPLREGRLYQVEYAMEEAISHA.GTCLGILS
RYDSRTTIFSP..EGRLYQVEYAQEAISNA.GTAIGTILS
RYDSRTTIFSP..EGRLYQVEYAMEEAISHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEEAIGHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEEAIGNA.GSALGVLA
RYDSRTTIFSP..EGRLYQVEYALEAINNA.SITIGLIT
SYDSRTTIFSP..EGRLYQVEYALEAINHA.GVALGIVA
    
```

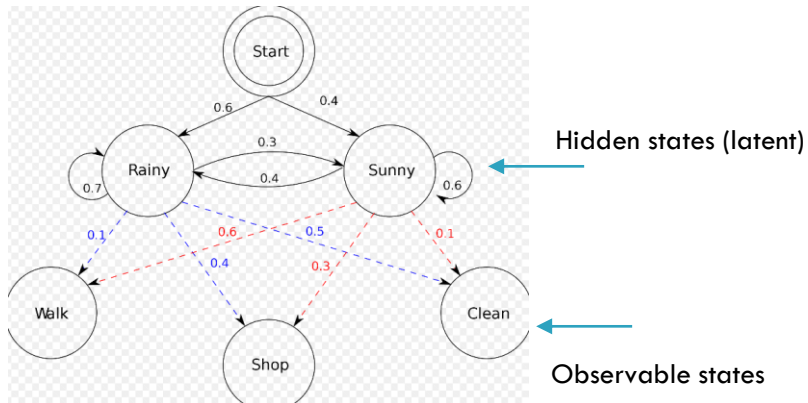
Training HMM



Optimal path  $\pi$  using DP



# Hidden Markov Model



WWWCCCSWW



SSRRRSSSS

# 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.    ==> FWD ... 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

HMMERS3/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
COMPO	m->m	m->i	m->d	i->m	i->i	d->m	d->d													
	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	2.96598	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
	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
	2.68618	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	3.02789	4.44770	4.32975	3.94537	3.37134	4.06203	4.67426	1.08311	3.77943	1.95821	3.30732	4.19130	4.49362	4.12214	3.97048	3.55690	3.32706	1.84575	5.26131	3.99926
	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													



1 m - -  
2 q - -  
3 i - -

# HMMer

## HMM profile indexing

```
hmmcompress clotting.hmm
```

```
[jwnam@biglab-master Session9]$ hmmcompress 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

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

```

# 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:          hmmsearch
# Version:          3.1b2 (February 2015)
# Pipeline mode:    SCAN
# Query file:       query.fa
# Target file:      clotting.hmm
# Option settings:  hmmsearch --domtblout test.tab clotting.hmm query.fa
# Current dir:      /home/jwnam/CLASS/Computational_biology/Session9
# Date:             Tue May 10 11:58:25 2016
# [ok]
```

query.tab



# parse\_hmmscan.py

```
#!/usr/bin/python                                     [jwnam@biglab-master Session9]$ python parse_hmmscan.py test.tab
import re                                             protname      len           domname       begin         end
import sys                                           FA7_HUMAN    466           clotting      27            121
import sys                                           FA7_HUMAN    466           clotting      204           454

filename = sys.argv[1]

print 'protname\tlen\tlname\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
```