

Basics of Markov chains

Samuel S. Shepard

Outline

- Discuss basic Markov models.
- Discuss a few applications.

Markov Chains

- Markov models are the basis for many gene prediction programs such as GeneMark.
 - ♦ GeneMark uses Hidden Markov models.
 - ♦ We developed a sequence prediction algorithm based on Markov chains called BAMM.
- Can apply to any sequence of information: nucleotide, amino acid, etc.

Markov models 101

- Markov models can be used to both *generate & classify* sequence data.
- The sequence frequency information must be analyzed first, then it can be used.
- Let's get a feel for Markov models with an analogy..

Fill in the blank...

(q r t e d u o s f m a y)

- i. th
- ii. gol
- iii. fluff
- iv. dinosau



Fill in the blank...

q r t e d u o s f m a y

- i. the
- ii. gol
- iii. fluff
- iv. dinosau



Fill in the blank...

q r t e d u o s f m a y

- i. the
- ii. gold
- iii. fluff
- iv. dinosau



Fill in the blank...

- i. the
- ii. gold
- iii. fluffy
- iv. dinosaur

q r t e d u o s f m a y



Fill in the blank...

- i. the
- ii. gold
- iii. fluffy
- iv. dinosaur

q r t e d u o s f m a y



Markov chain fundamentals

- The number of “letters” remembered by the Markov chain are known as its order.
- Markov chains can *generate* the next letter in the sequence based on the model frequencies.

Markov chain fundamentals

- Longer words like “dinosaur” were easier to guess than shorter ones like “gold” (could have been “golf”).
- Larger order Markov chains generally do *better* prediction.

Markov chains for Prediction

- Earlier you became human Markov models to generate words using your knowledge of *English*.
- *What if I only gave you a sequence of characters & wanted to know which language it was???*

Español or English?

tsnottearitthey said to
one another tets decid
e by lot who will get it thi
sh happened that the scr
ipture might be fulfilled
d that said they divided
my clothes among the
man and cast lots for my
garments so this is what
he sold



idamosse dijeron una
otra seche mosuertes p
ara ver a quien le tocase
lo hicieron los soldados
se los sucedió para que se
cumpliera la escritura que
dicese repartieron entre
ellos mimantoy sobre mir
opa echaron suer

Español or English?

tsnottear it **they** said to
one another tets decid
e by lot who will get it thi
sh happened that the scr
ipture might be fulfilled
d that said **they** divided
my clothes among the
man d cast lots for my g
arment so **this** is what t
he sold



idamosse dijeron una sa
otro se che mossuertes p
ar avera quien le toca ya si
lo hicieron **los** soldados se
stos sucedi o para que se c
umpliera **la** escritura que
dicese repartieron entre
ellos mimantoy sobre mi
ropa echaron suer

Doing Prediction

- Frequent patterns (words) help you see the *language* or model classification.
- It's difficult to make sense of the sentences without knowing where to start reading.

Help with Reading Frame

tsnottear itthey said to one another Le
tsdecide by lot who will get it This happ
ened that the scripture might be fulfilled
that said They divided my clothes a
mong them and cast lots for my garment
So this is what the soldi

Training for the Unknown

- Suppose you **don't know** either language.
- How do you do prediction without learning the meaning of every word in each language?

...beschlossensiediesesuntergewandwollen...

Training a Model

- You'd read lots of books in each language & learn the frequent words!



Example Training

- BAMM project used **6 million** nucleotides of exons and introns each.
- **3 million** bases are used to test prediction.

Markov chain types

- *Inhomogeneous* Markov models can “see” multiple reading frames.
 - ◆ Helps detect coding sequences.
 - ◆ More accurate.
- *Homogeneous* Markov chains don't care.

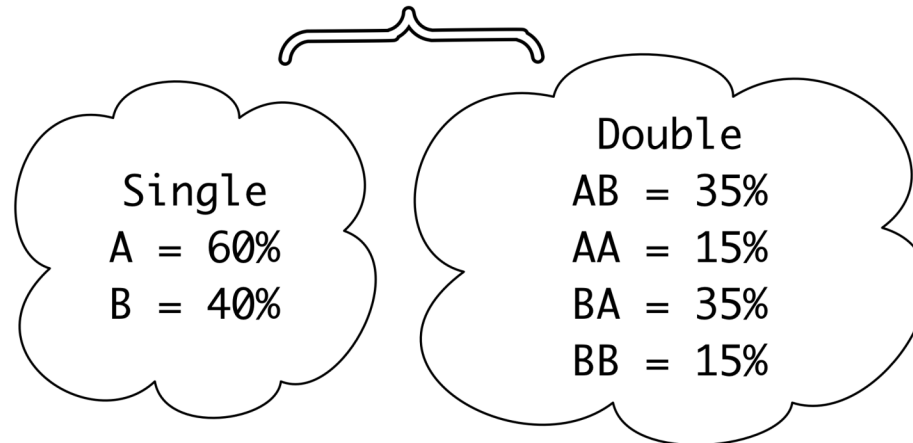
How does it work?

- Suppose we have been reading a lot of naturally occurring sequences represented by the alphabet $\{A,B\}$ and have come up with some frequencies.
- We can use this information for sequence generation (modeling) and classification (prediction).

Markov generation

- *Generating a sequence using a Markov model requires training first.*
- Frequency data is for *order 1* generation.

Order 1 model information (percent frequency)

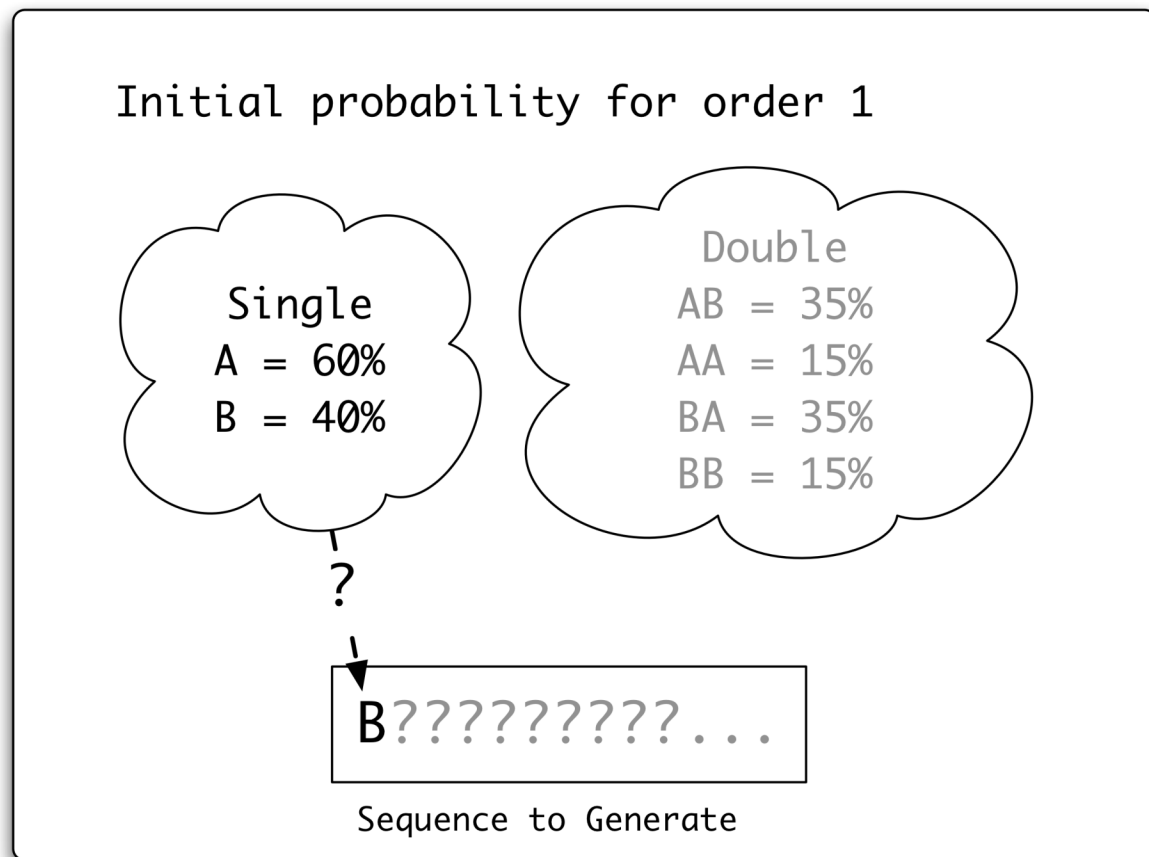


??????????...

Sequence to Generate

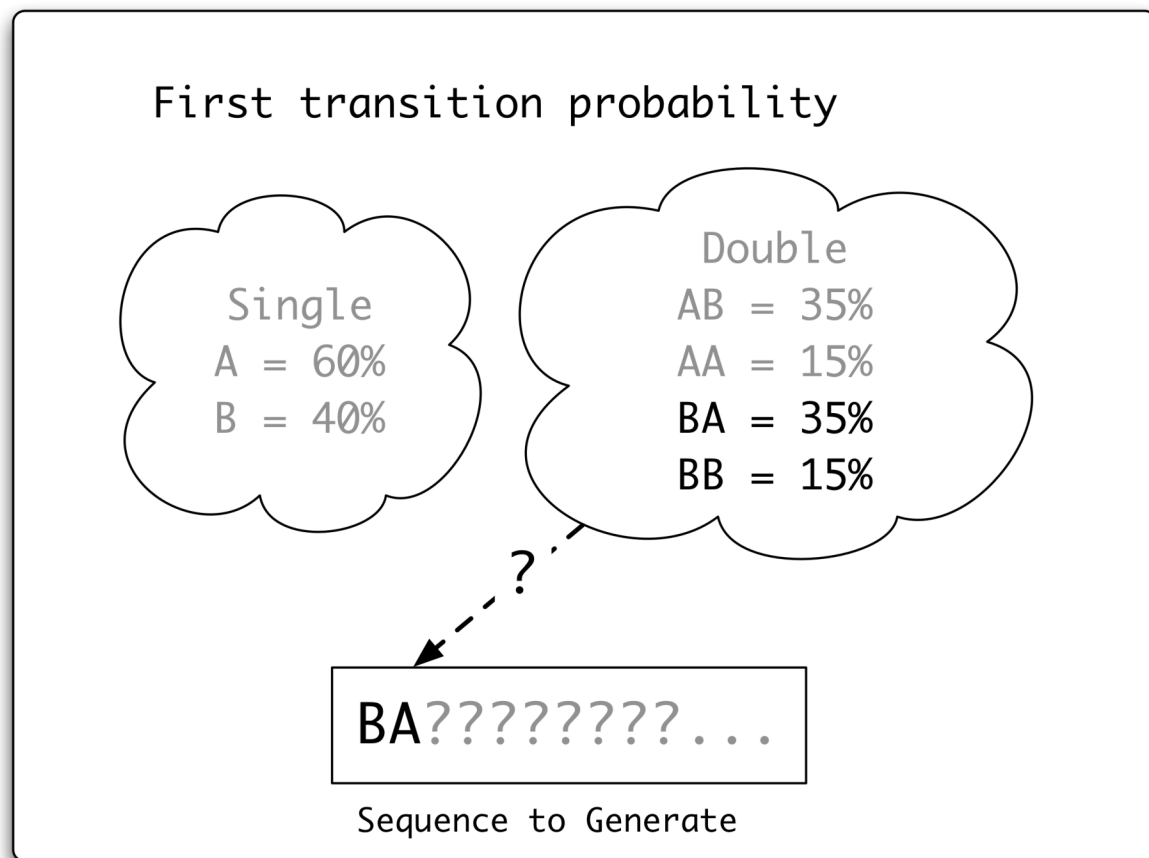
Markov generation

- To start the sequence, we use our initial probabilities.
- Generation is random, so each sequence can be unique.



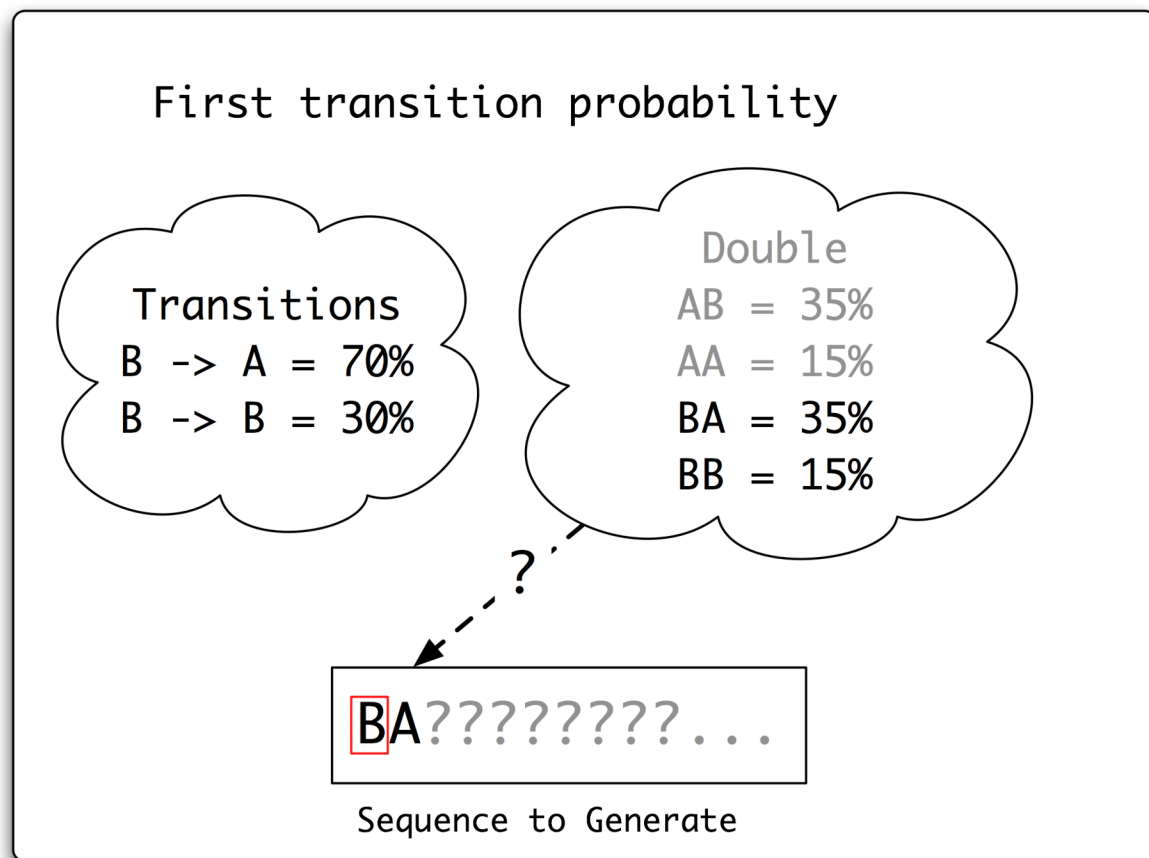
Markov generation

- Order-dependent transition probabilities are used to generate the next letter in the sequence.



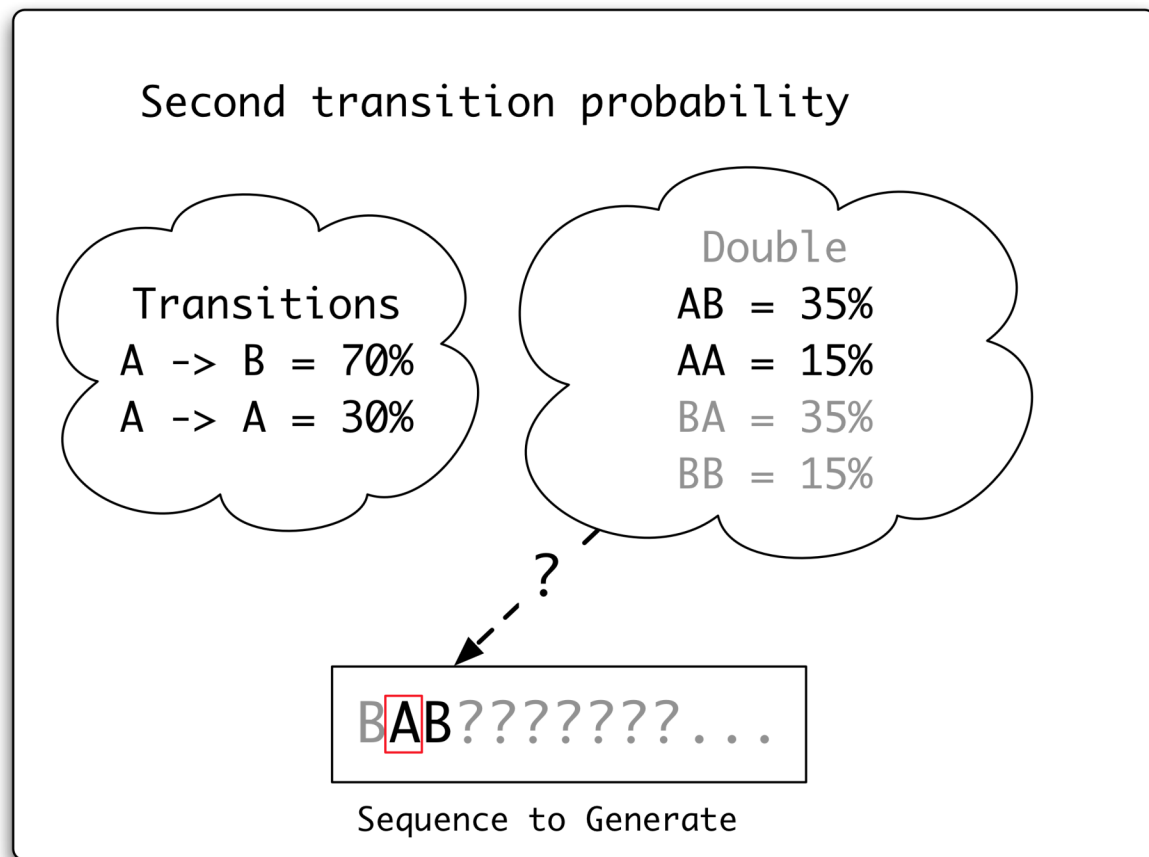
Markov generation

- Relative frequencies are used for the transition probabilities.
- These probabilities depend of the prefix [boxed], whose length is the order.



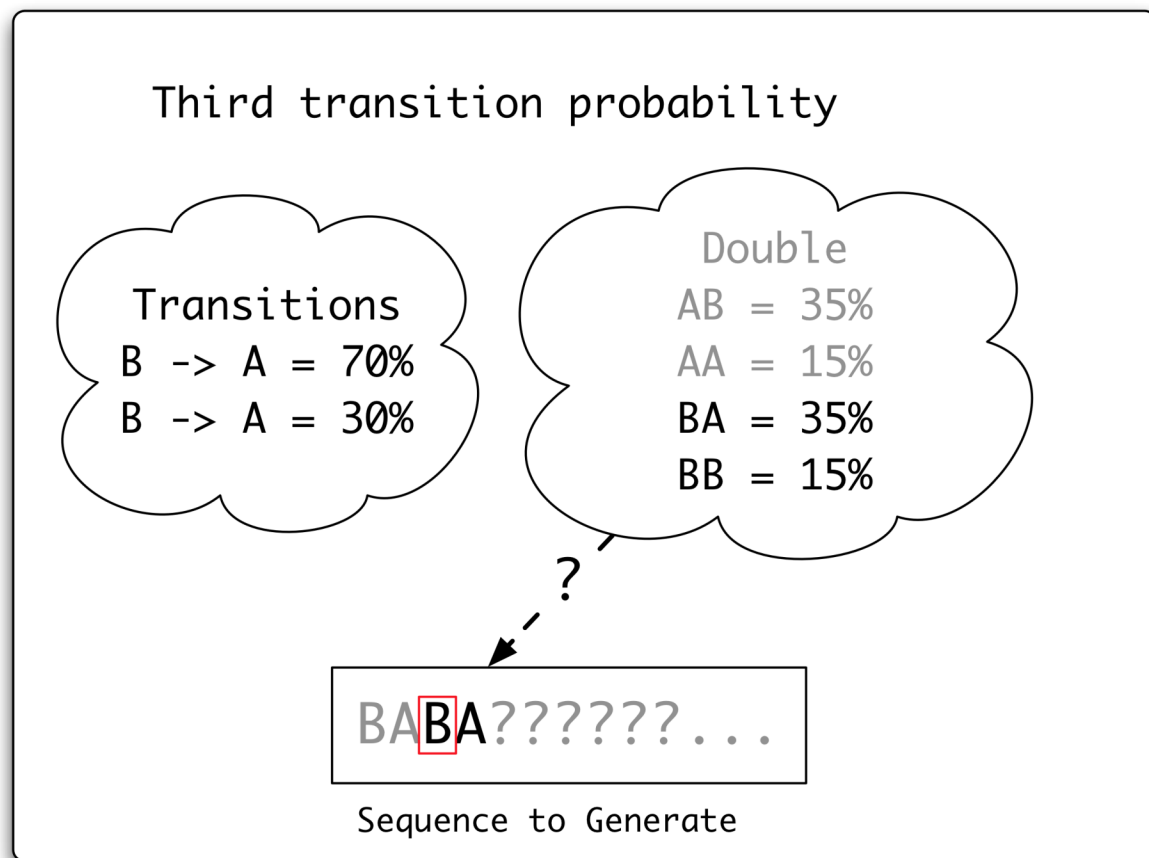
Markov generation

- Each new letter depends on the prefix, usually 0 to 5 bases for nucleotide sequences.



Markov generation

- The Markov chain algorithm continues as before until the desired number of letters is generated.



Markov generation

- While any permutation of the sequence is *possible*, not all sequences will be equally likely...

Single

A = 60%

B = 40%

Double

AB = 35%

AA = 15%

BA = 35%

BB = 15%

BABABBABABABA

Generated Sequence

Testing probability

- Suppose we have two sequences:
 - ♦ *BBBB*
 - ♦ *ABAB*

How likely is each sequence? Recall:

Single

A = 60%

B = 40%

Double

AA = 15% AB = 35%

BB = 15% BA = 35%

Probability of BBBB

- For the *BBBB* sequence, we get:
 - ♦ $B (0.40) \rightarrow B (0.30) \rightarrow B (0.30) \rightarrow B (0.30)$
 - ♦ Total probability = $(.4)(.3)(.3)(.3) = 0.0108$

What about *ABAB*?

Initial:
B = 40%

Transitions used:
B \rightarrow B = 30%

Probability of ABAB

- For the ABAB sequence, we get:
 - ♦ $A (0.60) \rightarrow B (0.70) \rightarrow A (0.70) \rightarrow B (0.70)$
 - ♦ Total probability = $(.6)(.7)(.7)(.7) = 0.2058$
- ABAB is a more probable.
 - ♦ $\text{Prob}(\text{ABAB}) = 0.2058 > 0.0108 = \text{Prob}(\text{BBBB})$

Initial:
A = 60%

Transitions used:
A \rightarrow B = 70%
B \rightarrow A = 70%

Models

- The total probability was determined by the initial & transitions probabilities. These probabilities characterize our model.
 - ♦ Let's call our previous example the "Ab model."
- Now consider a *null* model for uniformly random sequences:

Single

A = 50%

B = 50%

Transitions used:

AA = 25% AB = 25%

BB = 25% BA = 25%

Now under the *null* model

- $\text{Prob}(\text{BBBB} \mid \text{null}) = (0.5)^4 = 0.0625$
- $\text{Prob}(\text{ABAB} \mid \text{null}) = (0.5)^4 = 0.0625$
- Given sequence *ABAB*, what is the probability of the “Ab model” being used to generate it & not the null one?

A little likelihood

- Probability of “Ab model” given *ABAB* is about 77% versus the null model.

$$\begin{aligned} \text{Prob}(\text{“Ab model”} | ABAB) &= \frac{P(ABAB | \text{“Ab model”}) \cdot P(\text{“Ab model”})}{P(ABAB | \text{“Ab model”}) \cdot P(\text{“Ab model”}) + P(ABAB | \text{null}) \cdot P(\text{null})} \\ &= \frac{P(ABAB | \text{“Ab model”}) \cdot \frac{1}{2}}{P(ABAB | \text{“Ab model”}) \cdot \frac{1}{2} + P(ABAB | \text{null}) \cdot \frac{1}{2}} \\ &= \frac{P(ABAB | \text{“Ab model”})}{P(ABAB | \text{“Ab model”}) + P(ABAB | \text{null})} \\ &= \frac{0.2058}{0.2058 + 0.0625} \\ &= 77\% \end{aligned}$$

- Probability of “Ab model” given *BBBB* is less than 15% versus the null model.

Feeling the Odds

- Given some sequence S , what are the odds of that sequence being the “Ab model” versus null?

$$\begin{aligned}\text{Odds}(\text{sequence}) &= \frac{P(\text{model} = \text{“Ab model”} | \text{sequence})}{P(\text{model} = \text{null} | \text{sequence})} \\ &= \frac{P(S | \text{“Ab model”}) / P(S | \text{“Ab model”}) + P(S | \text{null})}{P(S | \text{null}) / P(S | \text{“Ab model”}) + P(S | \text{null})} \\ &= \frac{P(S | \text{“Ab model”})}{P(S | \text{null})}\end{aligned}$$

Feeling the Odds

- For the odds of *ABAB* we can see that:
 - ♦ $0.767/(1-0.767) = \underline{3.29} = 0.2058/0.0625$
- The odds of *BBBB* are: 0.172
- Normally, since Markov chains deal with very small probabilities, the chain is calculated in log-space.
- The score of a sequence being “Ab model” versus *null* is the log odds.
 - ♦ $\text{Score}(\textit{BBBB}) = \log(0.172) = -0.762$
 - ♦ $\text{Score}(\textit{ABAB}) = \log(3.29) = +0.517$

Markov chains in demand

- Markov chain log probabilities (or log odds) can be used by themselves or as part of more complicated prediction algorithms.
 - ◆ Hidden Markov model
 - ◆ Support vector machines (BAMM)

Binary-abstraction Markov model

“G” or not “G”, that is the question:

Binary-abstraction
process.

{ AGCTGTAATGTG . .
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
0 1 0 0 1 0 0 0 0 1 0 1 . .

The abstraction rule.

1 if G
0 otherwise

Markov Chain
Training/Testing

Abstraction Rule

- Abstraction rules indicate how to reduce nucleotide information into a binary code.
- Abstraction rules depend on the nucleotide word length.

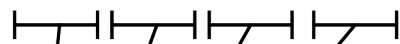
How many ways can I reduce nucleotide information?

Word Length	# Words	# Abstraction Rules
1	4	16
2	16	65,536
3	64	1.84×10^{19}
4	256	1.16×10^{77}

Nucleotides Words of Length 3

Binary-abstracted (BA3) Markov Model

AGCTGTAATGTG..



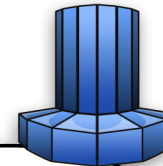
10010101001010110..

Markov Chain
Training/Testing

"GC-richness" Abstraction Rule

1 if $G+C \geq 2$
0 if $G+C < 2$

|—| = window of 3



* $G+C$ means "G or C"

Binary–abstraction Markov model

- Binary–abstraction Markov models allow one to analyze longer nucleotide sequence words by reducing the information analyzed.
- Analogous to replacing all articles in a sentence with ‘A’, verbs with ‘V’, and nouns with ‘N’, except in our case one must find what to replace first! Units of meaning are not obvious.

Profile HMMs

- Uses protein multiple sequence alignments to build an HMM profile of related proteins.
- The profile can be used to search for remote protein homologues within databases.

HMM Modeler

- Customizable profile HMM tool for remote homologue identification.
 - ♦ Implemented as a Chimera plug-in.
 - ♦ Joint effort of the Salzburg University of Applied Sciences with Salzburg University
- Astral Protein Database has protein sequences with less than 40% identity.
- SCOP protein families are grouped by structure.

Sample Alignment

Protein ID

Insertion

Match Column

Deletion

MSA

dla6ma_1a6m.pdb

-----VLSEGEWQLVLHVWAK-VEA-----

-----OVAGHGQDILIRLFKSHPETLEKFD

dlasha_1ash.pdb

-----ANKTRELCMKSLAH-AKVD-----

-----TSNARQDGIDLYKHMFEENYPLRKYF

dlb0ba_1b0b.pdb

-----LSAAQKDNVKSWSAK-ASA-----

-----AWGTAGPEFFMALFDAHDDVFAKFS

d1cg5b_1cg5.pdb

-----VKLSEDOEHYIKGVWVD-V-----

-----DIKQITAKALERVVFVVPWTTTLFS

d1cgxa1_1cgx.pdb

-----MLTQKTKDIVKATAPV-LAE-----

-----HYDIKCFYQRMFEAHPELKNVFN

dlecaa_1eca.pdb

-----LSADQISTVQASFDK-VKG-----

-----DPVGILYAVFKADPSIMAKFT

d1ew6a_1ew6.pdb

-----GFKQDIAT-IRG-----

-----DLRTYAQDIFLAFLNKYPDERRYF

d1gcva_1gcv.pdb

-----AFTACEKQTIGKIAQV-LAK-----

-----SPEAYGAECALARLFVTHPGSKSYF

d1gcwb_1gcw.pdb

-----VHWTQEEERDEISKTFQG-T-----

-----DMKTVVTTQALDRMFKVYPWNTNRYF

d1h97a_1h97.pdb

-----TLTKHEQDILLKELGP-HVDTP-----

-----AHIVETGLGAYHALFTAHPQYISHFS

d1hlba_1h1b.pdb

--GGT-----LAIQA-----QG-----

-----DLTLAQKKIVRKTWHQ-LMR-----

d1irda_1ird.pdb

-----VLSPADKTNVKAAGWK-VGA-----

-----HAGEYGAELERMFLSFPTTKTYF

d1it2a_1it2.pdb

-----PIIDQGPL-----

-----PTLTDGDKKAIKNIWPK-IYK-----

d1itha_1ith.pdb

-----GLTAAQIKAIQDHWFLNIKG-----

-----CLQAAADSIFFKYLTAYPGDLAFH

d1j17a_1j17.pdb

-----GLSAAQRQVVASTWKD-IAGAD-----

-----NGAGVGKECLSKFISAHPEMAAVF

d1mbaa_1mba.pdb

-----SLSAAEADLAGKSWAP-VFA-----

-----NKNANGLDFLVALFEKFPDSANFF

d1or4a_1or4.pdb

ET-----A-YFSDSNGQKRNRIQLTNKHA--DVKKQLKM---

-----VRLGDAELYVLEQLQPL-IQE-----

d1qlfa_1qlf.pdb

-----RPESELIRQSWRV-VSR-----

-----SPLEHGTVLFLARLFALEPSLLPLF

d1tu9a_1tu9.pdb

-----NAADRVMQSYGR-CCA-----

-----S-TGFFDDFYRHFCLASSPQIRAKF

d1urva_1urv.pdb

-----ELSEAERKAVQAMWAR-LYA-----

-----NSEDVGVAIVLVRFFVNFPSAKQYF

d1x9fa_1x9f.pdb

-----DCCSYEDRREIRHIWDD-VWSSSF--

-----TDRRVAIVRAVFDLKFHYPTSKALF

d1x9fb_1x9f.pdb

-----K--KQCGVLEGLKVKSEWGR-AYGS----

-----GHDREAFSQAIVRATFAQVPESRSLF

d1x9fc_1x9f.pdb

-----HEHCCSEEDHRIVQKQWDI-LWRDTES-

-----SKIKIGFGRLLLTAKLADIPEVNDLF

d2gdma_2gdm.pdb

-----GALTESQAALVKSSWEE-FNA-----

-----NIPKHTHRFFILVLEIAPAKDLFS

d2h8fb1_2h8f.pdb

-----VEWTDKERSIISDIFSH-M-----

-----DYDDIGPKALSRCLIVYPWTQRHFS

d3sdha_3sdh.pdb

-----SVYDAA-----AQLTADVKKDLRDSWKV-IGS-----

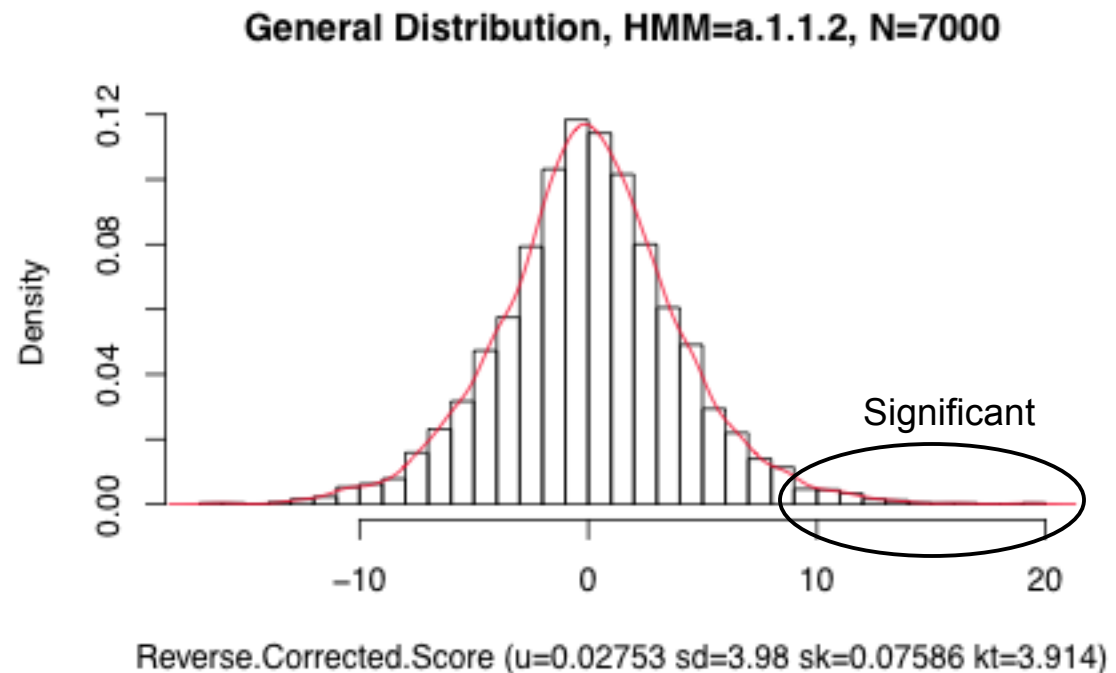
-----DKKNGVALMTTLFADNQETIGYF

Profile HMMs

- Match columns, deletions, and insertions are used to develop the profile HMM of the protein family.
- One can search protein databases using the profile, and based on the query score, filter for membership.

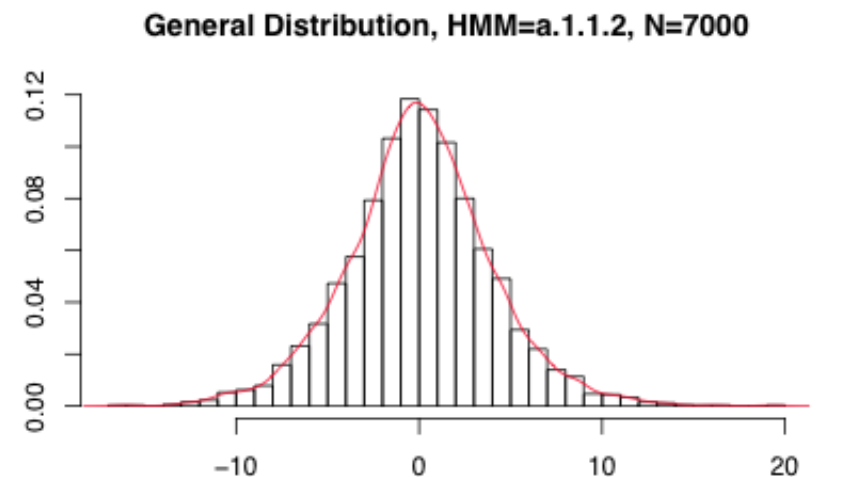
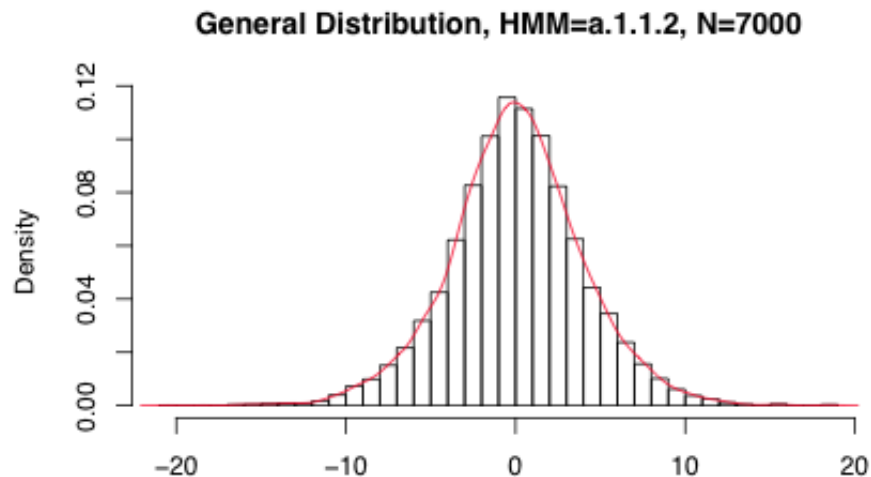
Evaluating remote homologues

- Scores are corrected for length bias.
- A *null* distribution is created of non-protein-family scores.
- Scores that exceed a threshold of significance, say greater than 95% could be counted.



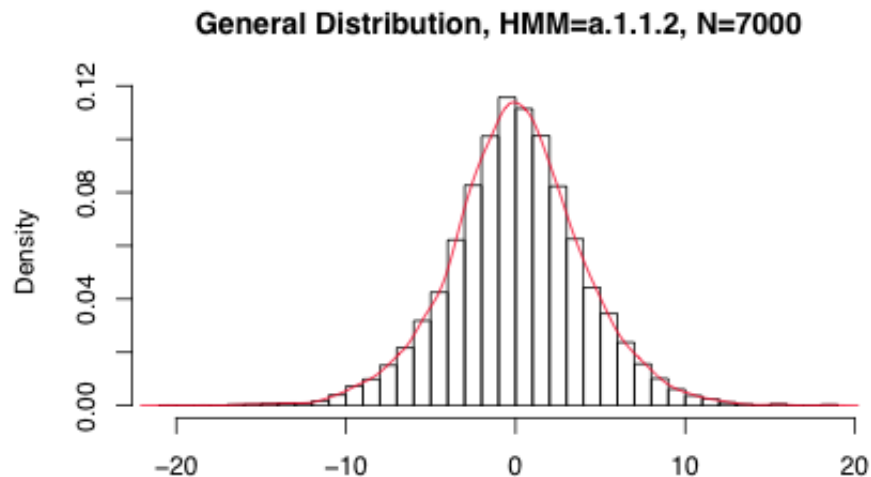
Simulating Proteins with Markov models

- Generated *null* distribution with Markov chain simulated proteins.

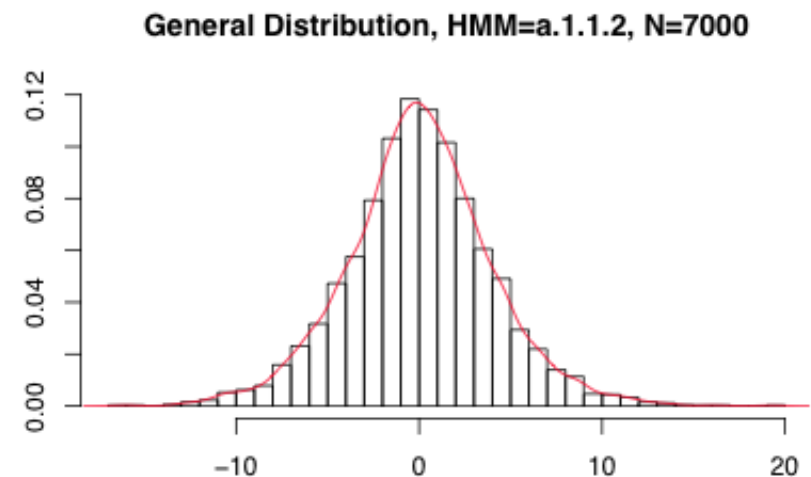


Simulating Proteins with Markov models

- Simulated proteins are generated from Astral database frequency information for Markov order 2.
- Which is the biological distribution?



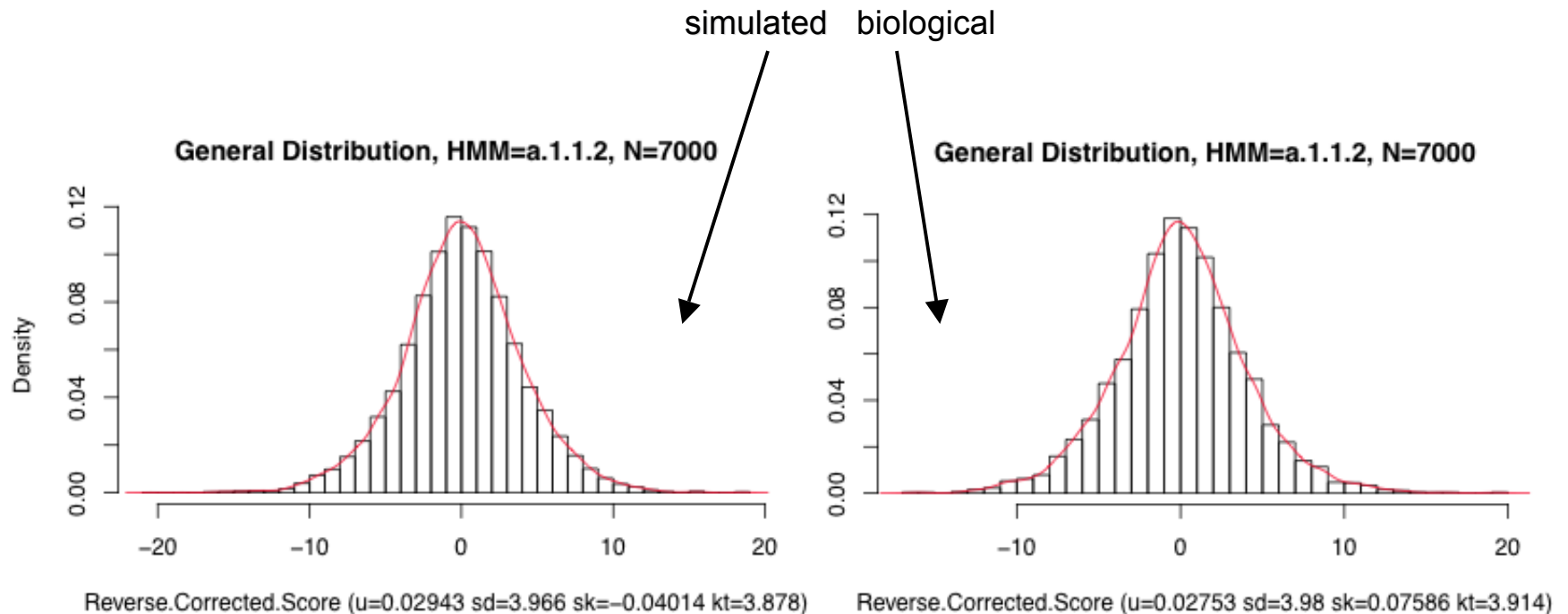
Reverse.Corrected.Score (u=0.02943 sd=3.966 sk=-0.04014 kt=3.878)



Reverse.Corrected.Score (u=0.02753 sd=3.98 sk=0.07586 kt=3.914)

Simulating Proteins with Markov models

- Simulated proteins can smooth the null distribution or reduce computation time.



In Conclusion

- Markov chains can be used for any sequence data.
- Useful in gene prediction, remote homologue identification, and much more.
- Can be used to generate AND discriminate sequence data.

Thank you for your attention.

Questions?

问题？

¿Preguntas?

Fragen?

вопросы?

質問か。