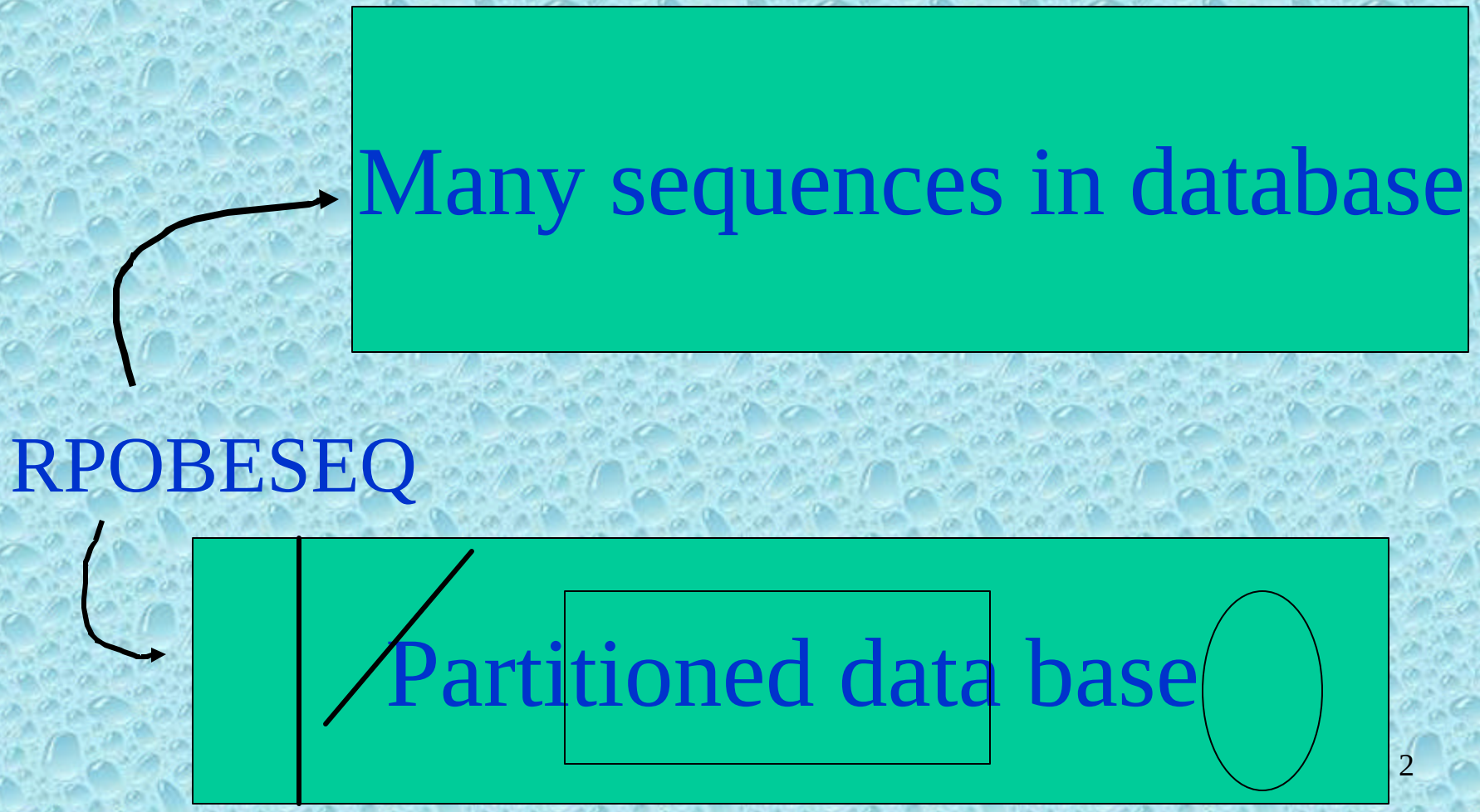
Hidden Markov Models in computational biology

Ron Elber

Computer Science

Cornell

Or: how to fish homolog sequences from a database

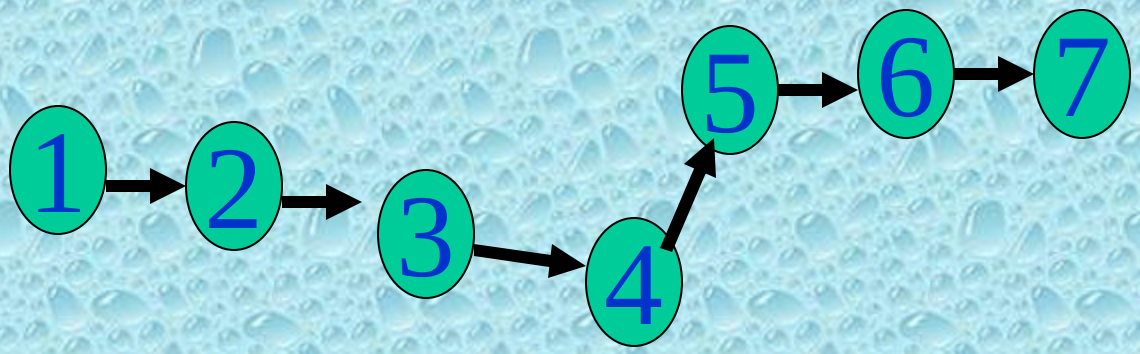


An accessible book on statistical models of sequence matching

R. Durbin, S. Eddy, A. Krogh and G. Mitchison, "**Biological sequence analysis: Probabilitic models of proteins and nucleic acids**", Cambridge, Cambridge 1998

What is a Markov model and why is it Hidden??

- A Markov process is used to describe time evolution. It depends only one step to the past p(i+1|i) – transition operator
- Sometimes we have a path but do not know the rules: the model is hidden.



HMM is a probabilistic model: A few words on probability

- Set of objects A_i (cities, amino acids):

- Probability of an object $P(A_i)$
- Joint probability $P(A_i, A_k)$
- Conditional probability $P(A_i | A_k)$
- Bayesian relationship

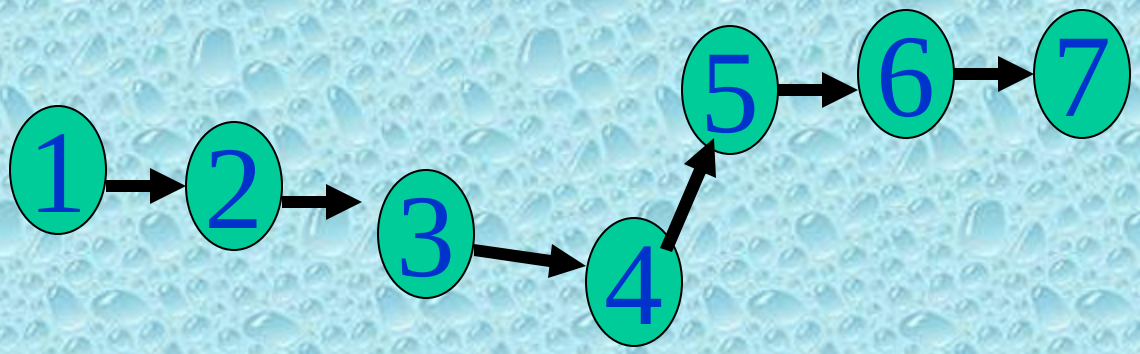
$$P(A_i, A_j) = P(A_i \mid A_j) P(A_j) = P(A_j \mid A_i) P(A_i)$$

Markov Models

$$P(1,2,3,4,5,6,7) \approx P(1|2)P(2,3,4,5,6,7)$$

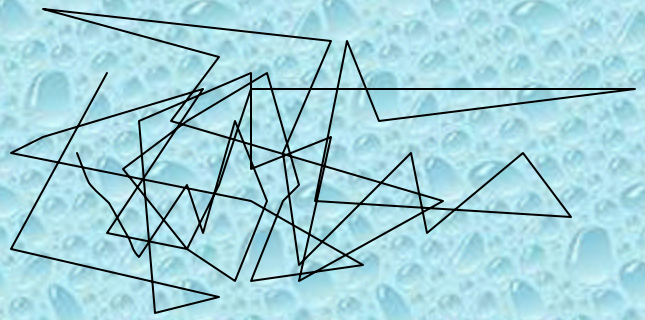
$$P(1|2)P(2,3,4,5,6,7) \approx P(1|2)P(2|3)...P(6|7)P(7)$$

In a Markov process the probability of item i+1 depends only on item i



Examples of Processes that are described by Markov models

- Walk under the influence
- Molecular diffusion in liquid
- Stock market
- Protein sequences



MKTTEYVAEILNELHNSAAYISNEEA

Is sequence scoring described by a Markov model?

```
Query: RPFAANGQSANYIPALGKVNDSQLGICVLEPDGTMIHAGDWNVSFTMQSISKVISFIAACMSRG
Sbjct: AVHYLNPESADLRALAKHLYDSYIKSFPLTKAKARAILTGKTTDKSPFVIYDMNSLMMGEDKIK
```

$$Score = \sum_i score(a_i, b_i)$$

Scoring sequence alignment is simpler than a Markov model

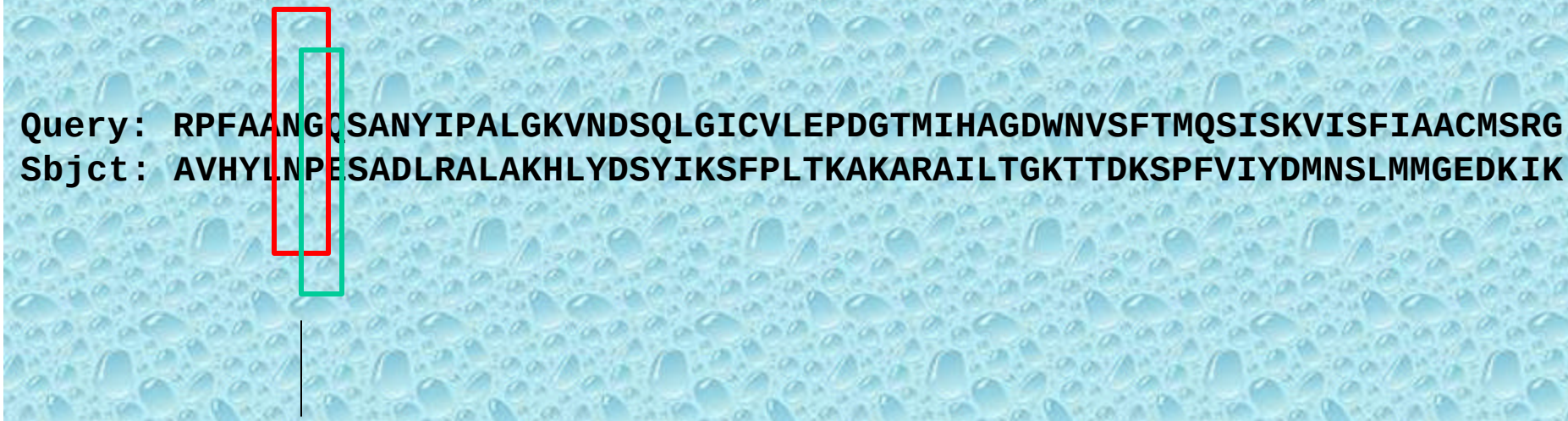
Scoring pair of sequences is independent on the order of the pairs

$$Score = \sum_i score(a_i, b_i)$$

In a Markov model we will expect scoring of the type:

$$Score = \sum_i score(a_{i+1}, b_{i+1} \mid a_i, b_i)$$

Simple Markov model for scoring sequence alignment



$$S = s(NN \mid GP) + s(GP \mid QE)$$

A Markov model of a single sequence

$$a_1 a_2 ... a_n \rightarrow P(a_1 \mid a_2) P(a_2 \mid a_3) ... P(a_n)$$

What is the Markov model of a "random" protein?

$$P(a_1, ..., a_n) \approx P(a_1) P(a_2) \cdots P(a_n) \approx \left(\frac{1}{20} \right)^n$$

A Markov model can be tailored to specific protein family

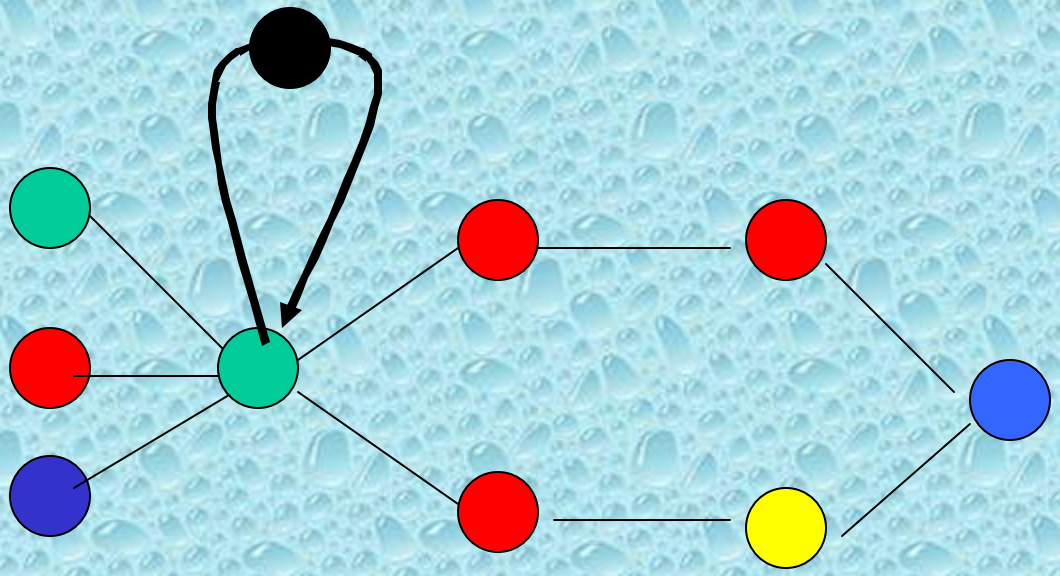
. . . . . HW . .

. . . . . HW . .

. . . . . HF . .

$$P(\,family) = \ldots\ldots d_H P(W \text{ or } F)\ldots$$

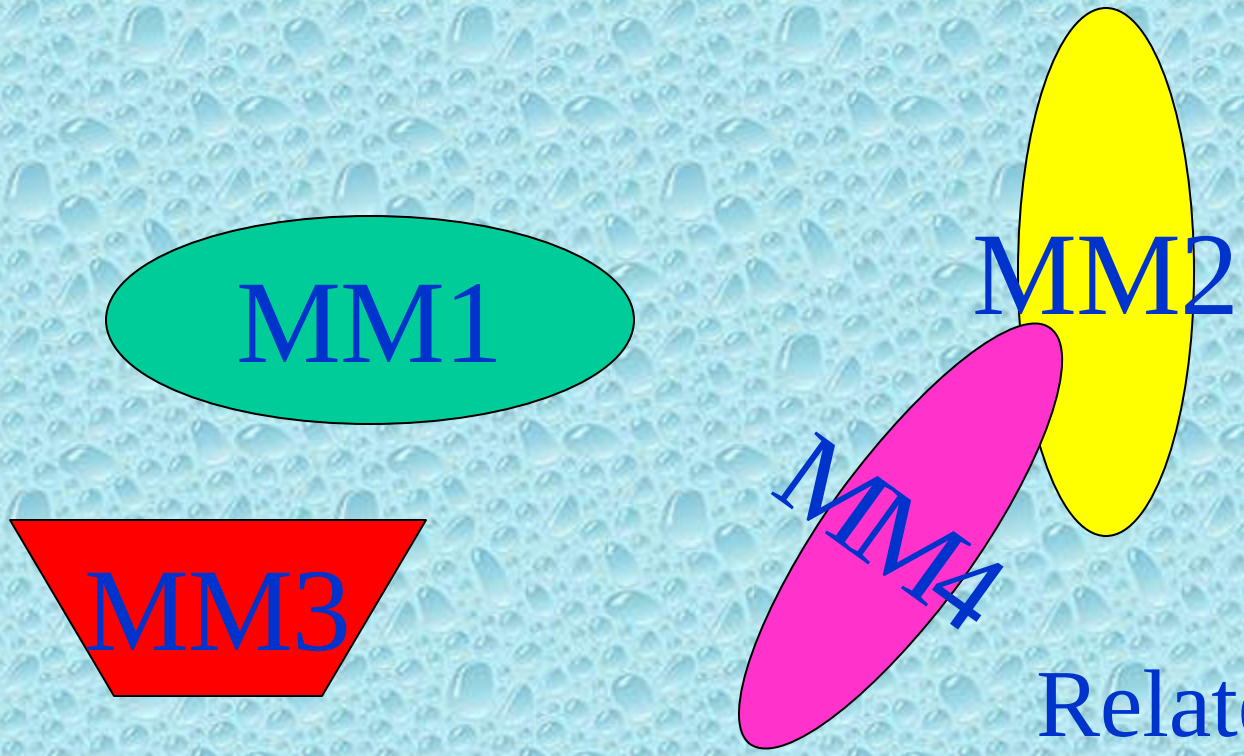
Pictorial view of a Markov model



Think on a Markov model for protein annotation as a **sequence generating machine**... Sequences are generated that are compatible with a given family.

For example, generate all sequences with histidine at position 64 and 93

The universe of sequences can be presented by HMM models



Related work by Golan Yona

How to use MM in sequence annotation

- Design Markov Models for known families
  - Use multiple sequence information
  - Estimate parameters for the Markov model
  - Generate compatible sequences, examine most probable sequence of the sequence generating machine. Most probable sequences can help identify family features

For annotation: check if a sequence is compatible with one of the MM models

Score against a Markov model

$$T = \sum \log\left[p(a_j \mid a_{j-1}) \right]$$

The sum is over multiple paths

Hidden Markov Models

- We have at hand a sequence with unknown annotation.
- We assume that the unknown sequence was generated by one of the Markov models of the known protein families.
- We wish to identify which of the (hidden) Markov Models is likely to generate the probe sequence.

A simple scoring scheme for a HMM model (profile)

- The protein family is denoted by M
- A sequence is denoted by S
- Are we interested in P(S|M) or P(M|S)?
- A simple probability model for P(S|M)

$$P(S \mid M) \cong \prod p_i(a)$$

- Yield a site-dependent scoring function

$$T = \sum \log\left(\frac{p_i(a)}{q(a)}\right) \equiv \sum t_i$$

- Is the general HMM score, can be written as a sum?

HMM for secondary structure

It is also possible to generate HMM for secondary structure prediction (based on correlation between 10-20 amino acids)

$P(sequence|helix)$ or $P(sequence|beta-sheet)$

The design of a “profile” model

- Have multiple sequence alignment

A A C H G W N ...
A V C F G W Q ...
A A C F G Y Q ...
A G C H A W D ...
A A C H G W Q ...

- Estimate probability of an amino acid at column i, p(a,i): problem with sampling

Estimating p(a,i)

- Use multiple sequence alignment to estimate frequencies $f(a,i)=n(a,i)/n(i)$
- In the simplest version the frequencies are set equal to probabilities $p(a,i)=f(a,i)$

Estimating p(a,i)

- Estimating p(a,i) is difficult
- Typically we will have sequences of about 10 family members
- For each site we have 20 possibilities (20 amino acids)
- If the probability is zero the score (-log of probability) is minus infinity
- **A major research in computer science: estimating model parameters from sparse data**

Estimating parameters of the model

Is a problem also for the simple profile model. Clearly a problem also for the more complex HMM models.

Pseudo counting is a simple workable procedure

- Have a reference zero order distribution (e.g. P(a) the probability of finding an amino acid independent of the site)
- Assume a set of points estimated from the reference distribution in addition to the multiple sequence alignment
- $P(a,i)=n(a,i)/n(i)+n(a)/n'$
- The more we have reference points the less specific is the model. A compromise between data and statistical accuracy

Summary HMM versus simple alignment

- HMM is a more complex model that includes information on multiple sequence alignment to parameterize a sequence-generating machine. As such it is expected to be more accurate than single sequence to a single sequence comparison.
- Depends on the availability of multiple sequences. A lot of work to generate the models...
- Extracting information from sparse data is a critical issue.