Google and Biosequence searches with Markov Chains

Nigel Buttimore

Trinity College Dublin

3 June 2010

Summary

- A brief history of Andrei Andreyevich Markov and his work
- Point Accepted Mutation matrix for protein sequences
- World Wide Web connectivity and its probability matrix
- Conclusions: Markov chain Biosequences and PageRank

A A Markov

- Born 14 June 1856 in Ryazan 160 km southeast of Moscow
- Markov graduated from St Petersburg University in 1878
- Markov chains: sequences of random variables in which the future variable is determined by the present variable but is independent of the way in which the present state arose from its predecessors.

- Markov processes in letter to Chuprov 15 January 1913
- Bicentenary of law of large numbers of J Bernoulli 1713
- Markov interested in poetry with studies of poetic style
- Probability of vowel for a preceding vowel or consonant
- Theory of stochastic processes by Kolmogorov in 1930s
- A A Markov died on 20 July 1922 in the then Petrograd

Probability Matrix

In a study of 20,000 successive letters Markov found that

- a vowel followed a vowel with probability 12.8%
- a vowel followed a consonant with chance 66.3%

when taken from Eugen Onegin by Pushkin (1799–1837)

$$\begin{pmatrix} 0.128 & 0.663 \\ 0.872 & 0.337 \end{pmatrix} \begin{pmatrix} 663 \\ 872 \end{pmatrix} = 1 \begin{pmatrix} 663 \\ 872 \end{pmatrix}$$

and so was derived the first example of a Markov matrix. In his correspondence with A. A. Chuprov of 1913 Markov says

> "To me this example seems instructive in many respects. I hope that no one has considered such an example up until now".

He did not foresee many applications of his ideas, such as the use in understanding biosequences of DNA and proteins; nor the critical rôle they play in the ranking of web pages.

DNA Base Mutations

Large Adenine and Guanine bases of DNA are purines (R). Small Thymine and Cytosine bases are pyrimidines (Y). Their proportions r and y change over a particular time T

$$\begin{bmatrix} r' \\ y' \end{bmatrix} = \begin{bmatrix} 0.91 & 0.11 \\ 0.09 & 0.89 \end{bmatrix} \begin{bmatrix} r \\ y \end{bmatrix}$$

such that after the period, $r = 30\%$ and $y = 70\%$ would become $r' = 35\%$ & $y' = 65\%$, then $r'' = 39\%$ & $y'' = 61\%$.

The ratio 55% :: 45% is fixed as $[\,11 \;\; 9\,]^{\mathrm{T}}$ is an eigenvector. All Markov matrices have an eigenvalue 1 and $|\lambda_i| \leq 1$, e.g.

$$\begin{bmatrix} 0.91 & 0.11 \\ 0.09 & 0.89 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = 0.8 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

indicating that the second eigenvalue is 0.8 so that the evolution of the numbers of large and small DNA bases is

$$\begin{bmatrix} r_n \\ y_n \end{bmatrix} = 5 \begin{bmatrix} 11 \\ 9 \end{bmatrix} + c \begin{bmatrix} -1 \\ 1 \end{bmatrix} 0.8^n$$

where c is a constant that depends on initial r & y values.

An example of a three by three probability matrix whose matrix elements are all strictly between zero and one and all of whose columns sum to unity, is the Markov matrix

$$\frac{1}{10}\begin{bmatrix} 6 & 1 & 1 \\ 3 & 7 & 1 \\ 1 & 2 & 8 \end{bmatrix} \begin{bmatrix} 4 \\ 7 \\ 9 \end{bmatrix} = 1 \begin{bmatrix} 4 \\ 7 \\ 9 \end{bmatrix}$$

where the leading eigenvector for eigenvalue 1 is shown. The leading eigenvector has strictly positive components. The order in size of the positive elements of this eigenvector play an important rôle in the PageRank algorithm of Google.

The eigenvectors for the matrix eigenvalues 0.6 and 0.5 are

$$\frac{1}{10}\begin{bmatrix} 6 & 1 & 1 \\ 3 & 7 & 1 \\ 1 & 2 & 8 \end{bmatrix}\begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} = \frac{3}{5}\begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

noting that eigenvector elements are not necessarily positive

$$\frac{1}{10}\begin{bmatrix} 6 & 1 & 1 \\ 3 & 7 & 1 \\ 1 & 2 & 8 \end{bmatrix}\begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}$$

in the case here of the non-leading eigenvalues less than 1.

PageRank

- The world's largest matrix computation: order 10 billion
- PageRank uses leading eigenvector of this large matrix
- Larry Page and Sergey Brin at Stanford University 1997
- Google's PageRank is recomputed reasonably frequently

- Does not involve any of actual content of Web pages
- Google lists query matches in order of their PageRank
- Randomly choose a link from one page to another one
- Can lead to dead ends at pages with no outgoing links — or to cycles around cliques of interconnected pages
- Choose random page on web a fraction 15% of the time
- Limiting probability for visiting a page is its PageRank

Idea of PageRank

- In essence, Google interprets a link from page A to page B as a vote by page A for page B.
- Google looks at more than the volume of votes, ie more than the links a page receives; ignores link farms.
- Google also analyses the web page that casts the vote.
- Votes cast by pages that are themselves "important" weigh more and help to make other pages "important".

Probability Matrix

Suppose G is the connectivity matrix of the web. Its element

$$g_{i\,j} \;=\; 1$$

if there is a hyperlink from page i to page j and 0 otherwise. G is huge as n is the total number of accessible web pages.

$$c_j \;=\; \sum_{i=1}^{n} g_{i\,j}$$

Quantities c_j are the column sums of the $n \times n$ matrix G and represent the number of links leading to the j-th page.

The transition probability matrix of the Markov chain used by Google is the $n \times n$ matrix A whose elements are

$$a_{ij} = \frac{0.85\, g_{ij}}{c_j} + \frac{0.15}{n}$$

The matrix elements are all strictly between zero and one. The columns of the matrix A all sum to one and therefore

$$A\, x = x$$

by a theorem of Perron and Frobenius indicating that

the largest eigenvalue of A is equal to one and that the dominant eigenvector is unique to within a scaling factor. After a Google search is performed the results of the search are ordered according to the elements of the leading eigenvector x forming the PageRank.

For more information see the article by Cleve Moler at

`www.mathworks.com/company/newsletters/news_notes/clevescorner/oct02_cleve.html`

on "The World's Largest Matrix Computation" from The MathWorks, MATLAB News and Notes — October 2002.

Related Web Links

mathworld.wolfram.com/MarkovMatrix.html

www.numbertheory.org/courses/MP274/markov.pdf

en.wikipedia.org/wiki/Markov_chain

www.statslab.cam.ac.uk/~mcmc/

www.ncbi.nlm.nih.gov/Education/blasttutorial.html

Conclusions

- Markov chains relate to processes with briefest history
- PAM probability matrix critical for biosequence seaches
- Connectivity matrix of web links is now 10 bn by 10 bn
- Google's PageRank is the state vector of a Markov chain