MEDVIX PUBLICATIONS

Journal of Medicine Care and Health Review
Research Article | Open Access | ISSN 3065-1719 | Volume 01 | Issue 03

The Chains of the Clusters of Latent States in DNA Sequencing

Orchidea Maria Lecian[1*]

[1]Sapienza University of Rome, Rome, Italy.

Received date: 23 September 2024; **Accepted date:** 10 October 2024; **Published date:** 19 October 2024

Corresponding Author: Orchidea Maria Lecian, Sapienza University of Rome, Rome, Italy.

Citation: Orchidea Maria Lecian. Sapienza University of Rome, Rome, Italy. Journal of Medicine Care and Health Review 1(3). https://doi.org/10.61615/JMCHR/2024/OCT027141019

Abstract

The possibilities to write Hidden Markov models from the cluster of latent states of DNA sequences are newly analytically investigated. The originating chains are studied. The sensitivity of the models to small perturbations is newly analytically proven. The comparison between the proposed model and those appearing in the literature is requested to be performed after the comparison of the distances of the corresponding graphs (on the opportune manifold) and of their differential, which is necessitated to compare the fundamental matrices of the originating chains as far as the Kolmogorov backward equations and the Kolmogorov forward equations are concerned.

Keywords: Latent states; clusters; DNA sequencing; amino acid substitution; graph theory; chains.

Introduction

The originating chains of the Clusters of latent states in DNA sequencing are studied in the present paper; the possibilities to write Hidden Markov Models out of the results are discriminated.

A Markov model of protein sequence was proposed in **[1]**.

The amino acid replacement models are considered: different patterns of amino- -acid replacements in different structural environments **[3]**, as explained in **[4]**; 'transmembrane-proteins' are examined to have a different replacement (of which the probability matrices are studied of the sequences). More specialized matrices are introduced in **[5]** and in **[6]**.

From **[7]**, some new timely difficulties in DNA sequencing are outlined; in particular, solving the epigenetic modifications for resolving the four DNA bases is instructed after a five-letter sequence workflow. Several samples are considered. Samples are divided into strands. The aim is to solve both genetic and epigenetics on the same 'read'.

From **[8]**, further aspects of genetic applications and epigenetic ones were integrated.

As described form **[9]**, the Hidden Markov Models gather interconnected states. Each state is discriminated after two kinds of parameters, ie. the symbol-emission probabilities and the state-transition probabilities. The symbol-emission probabilities represent the probability that each possible symbol is emitted out from a state. The state-transition probability expresses the transition to move from one state to another state.

A sequence is generated, which consists of states and probabilities to move from one state to another state.

The state sequence is proven to be a first-order Markov chain.

The prescriptions summed in **[10]** are here followed; in particular, the models of different events, which have an event-related dependence, are here developed: the differences between the two-states model and the n-states models, with $n > 2$, are outlined.

In the present paper, the possibilities to issue a chain originating a Hidden Markov Model from the cluster of the latent states are examined as far as the originating chain(s) of the corresponding Hidden Markov Model is concerned; the comparison with non-Markovian aspects is brought.

The results analytically prove the setting of **[11]**, which is further developed in **[12]**. In particular, from **[13]**, the property that the construction of the models is very sensitive to small perturbations in the data is newly proven, and the peculiarity that it is often inapplicable if the true distance is not small is newly delineated analytically.

The paper is organized as follows.

In Section 2, chain models of amino-acid substitution are discussed.

In Section 3, the possibility of writing analytically Hidden Markov Models from the cluster of latent states of the complete model is newly examined.

In Section 4, a comparison with other possible Hidden Markov Models is newly analytically brought.

In Section 5, the prospective studies are introduced.

2. About Chain Models of Amino-Acid Substitution

Form **[5]**, a model of amino-acid substitution in proteins is studied; more in detail, the general case is developed according to several analyses: i.e. both the Markovian approach and the non-Markovian developments are considered. The fundamental matrix of the chain $\hat{Q}$ is taken as $\hat{Q} = \hat{U}\,\hat{\boldsymbol{L}}\,\hat{U}^{-1}$, where $\hat{\boldsymbol{L}}$ is a diagonal matrix whose entries $\boldsymbol{l_{ii}}$, $\boldsymbol{i} = 1, 2, ..., 20$ are written as $\boldsymbol{l_{ii}} \equiv \boldsymbol{\lambda_i}$, being $\boldsymbol{\lambda_i}$ the eigenvalues of the fundamental matrix.

The probability matrix per unit time interval $\hat{M}$ is set. The matrix $\hat{}(U)$ is formed from the eigenvectors of the matrix $\hat{M}$.

The probability matrix $\hat{P}$ from the chain $\hat{Q}$ is written from the entries u_{ij} of $\hat{U}$ as

$$P_{ij} = \sum_k u_{ik} u_{kj}^{-1} e^{\lambda_k t}. \quad \textbf{(1)}$$

The matrix $\hat{R}$ is the matrix of the relative rate of substitution.

The matrix $\hat{A}$ is the mutation matrix. The frequency π^A of aminoacid i in the data is displayed in Table 22 of **[14]**.

2.1 The Memory-Less Markov Chain

The memory-less process is implemented from the elements in the above after relating the matrices $\hat{A}$ and $\hat{R}$ ·

In the Markovian case, the matrix $\hat{R}$ is the identity matrix $\hat{I}$.

Accordingly, the matrix $\hat{M}$ is specified as

$$m_{ij} = \delta/19, \quad i \neq j, \quad \textbf{(2a)}$$
$$m_{ij} = 1 - \delta, \quad i = j. \quad \textbf{(2b)}$$

The matrices A and R are related.

3. The Clusters of Latent States and the Possible Originating Chains

It is the purpose of the present Section to prove that a Hidden Markov Model of two-state amino-acid substitution as inspired by the interrogation in **[7]**, where selected items are considered only must descend from a chain different from that of **[5]**, independently on whether the chain is chosen as Markovian or as non-Markovian.

To this aim, one studies the properties of the matrix $\hat{M}$, and, in particular, of its eigenvalues μ_i to construct the eigenvectors to compose the matrix $\hat{U}$ from imposing $det[\hat{M} - \mu\hat{I}] = 0$. $\hat{}$I being the identity matrix.

In the case of a two-state model, the two eigenvalues of $\hat{M}$are $\mu_1 = 1$ and $\mu_2 = 1 - 2\delta$.

In the case of a Hidden Markov Model constating more than two states, the eigenvalue $\mu_j = 1, j = 1, 2, ...$ does not occur (proven by induction).

This way, the representation of the probability matrix according to $o(\delta)$ is comparably different in the case of the two-states Hidden Markov Model and in those of a different (greater) number of states.

This way, it is proven that the chain from which the two-state model looked for in **[7]** does not originate from the chain analyzed in **[5]**, independently of the choice of the process, i.e. either memory-less or memory.

This result analytically proves the findings of **[11]**, which are further developed in **[12]** and in **[13]**. In particular, the dependence of the model on small perturbations is here newly proven analytically.

4. Comparison with Other Possible Hidden Markov Models

The peculiarity of the two-state model is here outlined to be a unique one among the possible n-state Hidden Markov Models presented, i.e. in **[15,16,11]**,

The reversible Markov process model was schematized for nucleotide sequence analysis in **[16]**.

In **[15]**, a general reversible process model of Markov Process Models of nucleotide substitution is provided according to a 4×4 fundamental matrix of a Markov chain proposed from **[16]**; in this case, the peculiarities of the fundamental matrix allow one to infer that one of the parameters Π_i (from which the parameters π_j which relate the matrices $\hat{}$A and R$\hat{}$ in **[5]** are generalized) is redundant.

More in general, the comparison of the models **[5,16,15]**, consists of comparing the different probability matrices as far as the originating chains are concerned. For comparing the matrices, the distances between the corresponding graphs have to be computed, i.e. a Kantorovich distance as one from **[18]**; nevertheless, it is necessary to compare the fundamental matrices of the originating chains, for which purpose the definition of the derivatives is necessary to write the Kolmogorov backward equation and the Kolmogorov forward equation. The definition of distances and of their differentials can be found in the very recent approach **[19]**. Within this approach, it is possible to compare the presented results with those of **[20]**. The complete methodology for the comparison of the models is constituted therefore as defining the corresponding graphs (on the opportune manifold endowed with metric), the comparison being possible after having introduced the concepts of the distances between the graphs and of their derivatives. Within this framework, it is possible to compare the construction here presented with other schematizations, i.e. such as **[20]**.

5. Prospective Studies

From **[9]**, the simulated alignment is able to induce dynamics programming algorithms for the correct sampling of the suboptimal multiple alignments according to both the probability and the Markov landscape, which is shaped after the free energy, where the Markov landscape is expressed as from a 'Boltzmann temperature' factor.

The need for further algorithms to be implemented was recently expressed in **[17]**. Applications in molecular diseases can be found in **[21]**. Applications can be found in carcinogenesis **[8]**. The paradigms for comparison of the different possible chains are those eluci dated as comparin the distances among the corresponding graphs, as well as the differentials.

References

1. M.O. Dayhoff, R.V. Eck, C.M. Park, Washington,D.C. (1972). A model of evolutionary change in proteins, in M.O. Dayhoff, editor, Atlas of protein sequence and structure. Biomedical Research Foundation. 5: 89-99.
2. C. Kosiol. (2006). Markov Models for Protein Sequence Evolution, PhD Thesis, Cambridge University.
3. J. Overington, D. Donnelly, M.S. Johnson, A. Sali, T.L. Blundell. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds, Protein Sci. 1(2): 216-226.
4. D.T. Jones, W.R. Taylor, and J.M. Thornton. (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8(3): 275-282.
5. J. Adachi and M. Hasegawa. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA, J. Mol. Evol. 42(4): 459-468.
6. N. Goldman, J.L. Thorne, and D.T. Jones. (1998). Assessing the impact of sec- ondary structure and solvent accessibility on protein evolution, Genetics. 149(1): 445-458.
7. J. Fuellgrabe, W.S. Gosal, P. Creed. (2023). Simultaneous sequencing of genetic and epigenetic bases in DNA, Nat. Biotechnol. 41(10): 1457-1464.
8. J.M. Zingg, P.A. Jones. (1997). Genetic and epigenetic aspects of DNA methylation on genome expression, evolution, mutation and carcinogenesis, Carcinogen- esis. 18(5): 869-882.
9. S.R. Eddy. (1995). Multiple alignment using hidden Markov models, Proc. Int. Conf. Intell. Syst. Mol. Biol. 3: 114-120.
10. P. Hougaard. (1999). Multi-state Models: A Review, Lifetime Data Analysis. 5(3): 239-264.
11. T. Gojobori, K. Ishii K, M. Nei. (1982). Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotides, J. Mol. Evol. 18: 414-422.
12. T. Gojobori, W.H. Li, D. Graur. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes, J. Mol. Evol. 18(5): 360-369.
13. Z. Yang. (1992). Variations of substitution rates and estimation of evolutionary distances of DNA sequences, PhD Thesis, Beijing Agricultural University, Beijing.
14. M.O. Dayhoff, R.M. Schwartz, B.C, Orcutt Washington DC. (1978). A model of evolutionary change in proteins, in: M.O. Dayhoff (ed), Atlas of protein sequence and structure. National Biomedical Research Foundation. 5(3): 345-352.
15. S. Tavare. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences, in: Lectures in mathematics in the life sciences. 17: 57.
16. Z. Yang. (1994). Estimating the Pattern of Nucleotide Substitution, J. Mol. Evol. 39(1): 105-111.
17. K. Senthamarai Kannan, S.D Jeniffer, Hidden Markov Modelling for Bi- ological Sequence, in: R. Tiwari, M.F: Pavone, R. Ravindranathan Nair (eds). (2023). Proceedings of International Conference on Computational Intelli- gence; Algorithms for Intelligent Systems. Springer, Singapore.
18. M. Bernardo, M. Bravetti, (2003) Performance measure sensitive congruences for Markovian process algebras, Theoretical Computer Science 290, 117.
19. T. Brugere, Z. Wan, Y. Wang, Distances for Markov Chains, and Their Differentiation.
20. J. Henderson, S. Salzberg, K.H. Fasman, (1997) Finding genes in DNA with a Hidden Markov Model, J. Comput. Biol. 4, 127.
21. E. Zuckerkandl, L. Pauling. (1962). Molecular disease, evolution, and genetic hetero- geneity, in M. Marsha and B. Pullman, editors, Horizons in Biochemistry. 189-225.