社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－BIO－16 (8)
2009／3／6

木の最適ラベリング問題とその進化系統樹への応用

柳橋 史成† 伊藤 公人†† 有村 博紀†

† 北海道大学 大学院情報科学研究科 〒060–0814 札幌市北区北 14 条西 9 丁目
†† 北海道大学 人獣共通感染症リサーチセンター 〒001–0020 札幌市北区北 20 条西 10 丁目
E-mail: †fumi.yanagihashi@gmail.com, ††itok@czc.hokudai.ac.jp

あらまし 本稿では、葉にラベルを持つ木に対し、枝におけるラベルの差異が最小となるように、内部頂点にラベルを割り当てる問題 (OTLAP) を考察する。全ての割り当てを試す自明なアルゴリズムを用いた場合、頂点数 n の木に m 種類のラベルを最適に割り当てるときの時間計算量は $O(m^n)$ であり，頂点数 n の指数時間を要する。そこで本稿では、入力の木に対して，多項式時間で最適ラベリングを計算する動的計画法アルゴリズム DPAO を与える。アルゴリズムの時間計算量は，$O(km^2n)$ 時間であり，木の頂点数 n に関して線形である．ここに，k は木の最大次数とする．また、提案手法をインフルエンザウイルスの進化系統樹における仮想的分類単位の最適ラベリング問題に応用する。

キーワード 進化系統樹，動的計画法，木の最適ラベル割り当て，ベイズ推定

Optimal Label Assignment Problem for Rooted Trees and Its Application to Phylogenetic Trees

Fumiaki YANAGIHASHI†, Kimihito ITO††, and Hiroki ARIMURA†

† Graduate School of IST, Hokkaido Univ., N14 W9, Sapporo 060–0814, Japan
†† Research Center for Zoonosis Control, Hokkaido Univ., N20 W10, Sapporo 001–0020, Japan
E-mail: †fumi.yanagihashi@gmail.com, ††itok@czc.hokudai.ac.jp

Abstract In this paper, we study the optimal tree label assignment problem(OTLAP), and present an efficient dynamic programming algorithm DPAO that solves the OTLAP in $O(km^2n)$ time for an input tree with maximum degree k and size n and a $m \times m$ cost matrix over a label alphabet of size m. We then apply our algorithm to the optimal labeling inference problem for the phylogenetic tree of influenza viruses.

Key words phylogenetic tree, dynamic programming, optimal tree label assignment, Bayesian inference

1. はじめに

近年，計算機の急速な発展により，大規模な木構造データを扱う機会が急増している．生命情報科学の分野では，大量の遺伝子配列から推定される進化系統樹が例として挙げられる．このような状況の中，大規模な木構造データを高速に処理する技術の開発は重要な課題である．

本研究では、木に対する最適ラベリング問題 (OTLAP) を扱う．OTLAP とは，葉にラベルを持つ木に対し、枝におけるラベルの差異が最小となるように、内部頂点にラベルを割り当てる問題である。木の頂点数を n とし，割り当てるラベルの種類数を m とすると，全ての割り当てを試す自明なアルゴリズムの時間計算量は $O(m^n)$ である。そこで本研究では，この問題に対する効率のよいアルゴリズムを開発することを目的とする．

本研究では、入力の木に対して、多項式時間で最適ラベル割り当てを計算する動的計画法アルゴリズム DPAO を与える。また，本手法をインフルエンザウイルスの遺伝子解析に応用し，進化系統樹と分離地域のデータから先祖配列の地理ラベルを推定する．そして，インフルエンザウイルスの地理的な広がり方の特徴を明らかにする．

木の最適ラベリング問題の関連研究として，進化系統樹を求める最節約法が挙げられる．しかし，最節約法は葉となる頂点の遺伝子配列から進化系統樹と内部頂点の遺伝子配列を同時に推定する問題を扱い，最適解を厳密に計算することが困難であることが知られている [6]．生命情報科学では，ペアワイズアラインメント等に動的計画法が応用されている．

本稿の構成は次のとおりである．2 節で最適ラベリング問題を導入し，3 節で動的計画法を用いた最適化アルゴリズムを示す．4 節でベイズ学習の枠組みに基づくラベル割り当てを考察し，5 節でインフルエンザウイルスの進化系統樹に関する実験結果を報告する．6 節でまとめる．

2. 準 備

本節では，基本的な概念を導入する．本稿にない用語については，[1], [3] 等を参照されたい．以下では，$\mathbb{R}$ と $\mathbb{R}_+$ で，それぞれ**実数全体の集合**と**非負実数全体の集合**を表す．正整数 $n \geq 1$

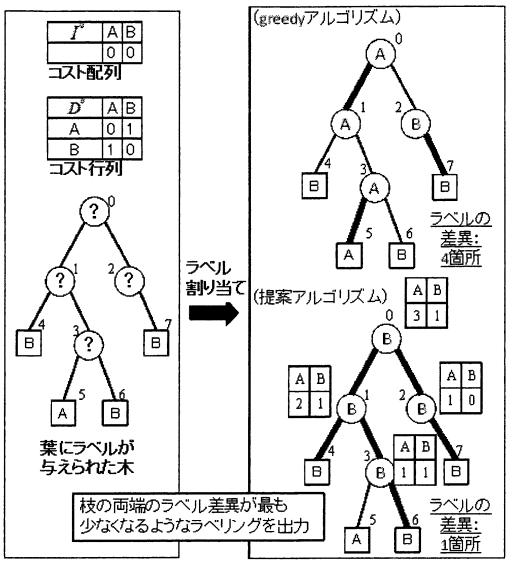

図 1　木の最適ラベリング問題

に対して，$n \times n$ 行列を $A = (a_{ij})$ と書く．

2.1 根付き木

本稿の対象は，図 1 に示したような**根付き木**（単に**木**と呼ぶ）$\mathcal{T} = (V, E, root)$ である．ここに，V は頂点集合であり，$E \subseteq V^2$ は**枝**の集合，$root \in V$ は，**根**である．親と子，葉，内部頂点などの用語は通常通り定義する [1] 本稿では，木 $\mathcal{T}$ に対して，頂点集合を $\mathcal{T}$ と書き，$\mathcal{T}$ の葉全体と内部頂点全体をそれぞれ $\mathcal{L}$ と $\mathcal{A}$ で表す．図 1 で 0〜3 番の頂点が $\mathcal{A}$ であり，4〜7 番の頂点が $\mathcal{L}$ である．これより，$\mathcal{T} = \mathcal{A} \cup \mathcal{L}$ である．

2.2 ラベル割り当てとコスト関数

$\Sigma = \{1, \ldots, m\} (m \geq 1)$ を**ラベル**の集合とする．未定義値を $\perp$ とおく．$\mathcal{T}$ への**ラベル割り当て**とは，部分関数 $\ell_0 : \mathcal{T} \to \Sigma \cup \{\perp\}$ である．$\mathcal{T}$ への**完全ラベル割り当て**は，頂点 x にラベル $l(x)$ を割り当てる関数 $\ell : \mathcal{T} \to \Sigma$ をいう．各頂点 $x \in \mathcal{T}$ に対して，$\ell(x) \in \Sigma$ を ℓ による**頂点 x のラベル**と呼ぶ．ラベル割り当て ℓ, ℓ' に対して，$\forall x \in \mathcal{T}, \ell(x) \in \Sigma \Rightarrow \ell(x) = \ell'(x)$ ならば，$\ell \subseteq \ell'$ と書く．

$\mathcal{T}$ 上の完全ラベル割り当ての**コスト関数**とは，任意の完全ラベル割り当て ℓ に対して，そのコスト $Cost(\ell) \in \mathbb{R}_+$ を割り当てる関数 $Cost$ であり，長さ m のコスト配列 $I = (I_i)$ と $m \times m$ のコスト行列 $D = (D_{ij})$ によって与えられる．ここに，I と D の要素は $\mathbb{R}_+$ の要素である．コスト行列は対称でなくても良い．

[例 1]　$\Sigma = \{1, \ldots, m\}$ $(m \geq 1)$ とおく．次のコスト配列 I^0 とコスト行列 D^0 で与えられるコスト関数を考える．$I^0_i = 0 (i = 1, ..., m)$ である．$i = j$ のとき $D^0_{ij} = 0$ であり，$i \neq j$ のとき $D^0_{ij} = 1$ である．これは，同一のラベル対にコスト 0 を割り当て，異なるラベル対にコスト 1 を割り当てるコスト関数である．図 2 に，ラベル集合サイズ $m = 4$ の場合の行列 D_0 を示す．

2.3 最適ラベリング問題

根付き木 $\mathcal{T} = \mathcal{A} \cup \mathcal{L}$ の最適ラベリング問題は，次のように定義される最適化問題である．

I^0	1	2	3	4
	0	0	0	0

D^0	1	2	3	4
1	0	1	1	1
2	1	0	1	1
3	1	1	0	1
4	1	1	1	0

図 2　ラベル集合サイズ $m = 4$ のコスト配列 I^0 とコスト行列 D^0

[定義 1]　**木に対する最適ラベリング問題**
(Optimal Tree Label Assignment Problem, OTLAP)
入力：
葉集合 $\mathcal{L}$ と内部頂点集合 $\mathcal{A}$ をもつサイズ $n \geq 0$ の根付き木 $\mathcal{T}$，サイズ $m \geq 1$ のラベル集合 $\mathcal{T} = \mathcal{A} \cup \mathcal{L}$，長さ m のコスト配列 I, $m \times m$ コスト行列 D, $\mathcal{T}$ の葉へのラベル割り当て ℓ_0.
問題：
$\mathcal{T}$ への完全ラベル割り当て $\ell : \mathcal{T} \to \Sigma$ で，$\ell_0 \subseteq \ell$ を満たし，D に関するコスト

$$Cost(\ell, I, D, \mathcal{T}) = I(\ell(root)) + \sum_{(x,y) \in E} D_{\ell(x)\ell(y)} \qquad (1)$$

を最小化するものを見つけよ．

3. 多項式時間アルゴリズム

本節では，最適ラベリング問題を，動的計画法を用いて効率よく解くアルゴリズム DPAO（Dynamic Programming Algorithm for OTLAP）を与える．

図 3 に，アルゴリズム DPAO の概要を示す．

アルゴリズムは，**DP 表**と呼ばれる，部分問題の最適コストを保存するための二次元配列 $BS : \mathcal{T} \times \Sigma \to \mathbb{R}_+$ をもつ．各要素 $BS[v][a]$ は，頂点 v にラベル $a \in \Sigma$ を割り当てたときの部分木 $\mathcal{T}(v)$ の最適コストである．

図 4 に示す手続き ComputeTable により木を葉から根まで上昇しながら DP 表を計算する．(1) 頂点 v が葉のとき，v にはラベル $l(v)$ が割り当てられているので，コスト行列 D の値から $BS[v][a] = D_{l(v)a}$ である．(2) 頂点 v が内部頂点のとき，v の子を $v_i (i = 1, 2, ...)$ とすると，v のコストは値 BS_i の総和 $BS[v][a] = \sum_i BS_i$ である．ここで，BS_i は，枝 (v, v_i) の距離と DP 表の値 $BS[v_i][b] (b \in \Sigma)$ の和の最小値であり，

$$BS_i = min\{D_{ab} + BS[v_i][b] \mid b \in \Sigma\}$$

である．

図 5 に示す手続き TraceBack は，木を根から葉に下降しながら，計算済みの DP 表の値から，各頂点 v に対して最適ラベル割り当て $BL[v]$ を次のように計算する．頂点 v が内部頂点のとき，DP 表 $BS[v][a]$ が最小となるようなラベル $a \in \Sigma$ を最適ラベルとして割り当てる．頂点 v が葉のとき，葉へのラベル割り当て l_0 を最適ラベルとして割り当てる．

[定理 2]　図 3 のアルゴリズム DPAO は，木に対する最適ラベリング問題を $O(km^2n)$ 時間で解く．ここに，k は入力木 $\mathcal{T}$ の最大次数であり，$m = |\Sigma|$ はラベル集合のサイズ，$n = |\mathcal{T}|$ は $\mathcal{T}$ の頂点数である．

(証明)　手続き ComputeTable は初めに，図 4 の 11 と 12 行目で，BS_i の値を $O(m)$ 時間で計算し，次に，BS_i を v の数だけ足し合わせて，$BS[v][a]$ を $O(k)$ 時間で計算する．さらに，図 4 の 9〜14 行目で，各ラベルについて $BS[v][a]$ を $O(m)$ 時間で計算する．以上の DP 表の計算を，n 個のノードについて再帰

Algorithm DPAO($\mathcal{T}, \Sigma, D, \ell_0$):
Input: 根付き木 $\mathcal{T} = (V, E, root)$, ラベル集合 $\Sigma = \{1, \ldots, m\}$, Σ 上のコスト行列 $D \in \mathbb{R}_+^{m \times m}$, $\mathcal{T}$ の葉へのラベル割り当て $\ell_0 : \mathcal{L} \to \Sigma \cup \{\perp\}$.
Output: 最適ラベル割り当て $BL : \mathcal{T} \to \Sigma$
大域変数: 最適スコアの二次元配列 $BS : \mathcal{T} \times \Sigma \to \mathbb{R}_+$ と,
最適ラベルの一次元配列 $BL : \mathcal{T} \to \Sigma$;

```
1: ComputeTable(root, w, D, Σ, BS);
2: TraceBack(root, ℓ0, BS, BL);
3: return BL;
```

図 3 最適ラベリング問題の多項式時間アルゴリズム．動的計画法を用いて，ボトムアップに解を計算する．

Procedure ComputeTable(v, w, D, Σ):
Input: 頂点 v;
Task: 木を葉から根まで上昇しながら，各頂点 v に対して，動的計画法の表 $BS[v] : \Sigma \to \mathbb{R}_+$ を計算する;

```
1: if v が葉である then
2:   for all ラベル a ∈ Σ do
3:     BS[v][a] = D_{l(v)a}
4:   end for
5: else
6:   for all v の子供 u do
7:     ComputeTable(u, w, D, Σ);
8:   end for
9:   for all ラベル a ∈ Σ do
10:    /* BS[v][a] を計算する． */
11:    BS_0 = min{ D_ab + BS[v_0][b] | b ∈ Σ };
12:    BS_1 = min{ D_ab + BS[v_1][b] | b ∈ Σ };
13:    BS[v][a] = BS_0 + BS_1;
14:  end for
15: end if
```

図 4 再帰手続き ComputeTable

Procedure TraceBack(v, ℓ_0, BS, BL):

```
1: if v が葉である then
2:   BL[v] := ℓ0(v);
3: else
4:   BL[v] = argmin_a{BS[v][a] | a ∈ Σ}
5:   for all v の子供 u do
6:     TraceBack(u, ℓ0, BS, BL);
7: end if
```

図 5 再帰手続き TraceBack

呼び出しで繰り返すので，手続き ComputeTable の総計算時間は，$O(km^2n)$ である．同様に，手続き TraceBack は各頂点ごとに $O(m)$ 時間で $BS[v][a]$ が最小値をとるラベル a を見つける．したがって，手続き TraceBack の総計算時間は $O(mn)$ である．以上より，アルゴリズム DPAO の総計算時間は，$O(km^2n)$ である． □

4. 進化系統樹の最適ラベリング問題への応用

本節では，ベイズ学習 [2] の枠組みにしたがって，進化系統樹の最適ラベル割り当てのための確率モデルを与える．**進化系統樹**を，頂点集合 $V = \mathcal{A} \cup \mathcal{L} = \{1, \ldots, n\}$ をもつ n 頂点で次数 2 の根付き木 $\mathcal{T}$ とする．内部頂点は推測された先祖配列に，葉は実際に観測された子孫配列にそれぞれ対応する．

本節では，$\mathcal{T}$ の頂点への**確率的なラベル割り当て** ℓ を考え，**ラベル割り当ての生成モデル** $p(X, Y)$ を，木 $\mathcal{T}$ を分解モデルにもつグラフィカルモデル [2], [3] として，以下のように与える．頂点のラベルを表す**確率変数**を $x_i \in \Sigma$ で表す．さらに，$\mathcal{A}$ と $\mathcal{L}$ に対するラベル割り当てを，それぞれ，組 $X = (x_1, x_2, \ldots, x_{n-l}) \in \Sigma^{n-l}$ と，組 $Y = (x_{n-l+1}, \ldots, x_n) \in \Sigma^l$ で表す．**初期生起確率** $q(x_0)$ は，根 $root$ がラベル $x \in \Sigma$ をもつ初期確率である．**置換確率** $p(y \mid x)$ は，親配列がラベル $x \in \Sigma$ をもつとき，これが子配列でラベル $y \in \Sigma$ に変化する条件付き確率である．実際には，$q(x_0)$ と $p(y \mid x)$ は，実験による経験確率等で求める．ここで，頂点 $i \in V$ の親を $\pi(i) \in V$ で表すと，i のラベルは x_i で，その親のラベルは $x_{\pi(i)}$ と書ける．

[定義 2] (進化系統樹のラベル割り当て確率) 木 $\mathcal{T}$ がラベル割り当て $Z = (X, Y)$ をもつ確率 $p(X, Y)$ を，次のような確率分布として与える．

$$p(X, Y) \;=\; q(x_0) \prod_{i=1}^{n} p(x_i \mid x_{\pi(i)}) \qquad (2)$$

我々の学習問題は，**事後確率最大化** (MAP, Maximum a Posteriori Probability) 基準による内部頂点へのラベル割り当て X の学習である．これは，葉へのラベル割り当て Y が与えられたとき，事後確率 $p(X \mid Y) \;=\; p(X, Y) \,/\, p(Y) \;=\; p(X, Y) \,/\, \sum_{X'} p(X', Y)$ を最大化する内部頂点へのラベル割り当て X を求める問題である．

[定理 3] 式 (2) で定義される確率分布 $p(Z)$ に関して，事後確率 $p(x \mid y)$ を最大化するラベル割り当て ℓ は，以下のように定義されるコスト配列 $I = (I_x)$ とコスト行列 $D = (D_{xy})$ に関する OTLAP 問題の最適解に一致する．

$$\begin{aligned} I_x &= -\log q(x), \qquad \text{for } x = 1, \ldots, m \\ D_{xy} &= -\log p(y \mid x), \quad \text{for } x, y = 1, \ldots, m \end{aligned}$$

(証明) 初めに，$I(x) = I_x = -\log q(x)$ および $D_{xy} = D_{xy} = -\log p(y \mid x)$ とおくと，次の導出が得られる．

$$\begin{aligned} \hat{X} &= \mathbf{argmax}_X \; p(X \mid Y) \\ &= \mathbf{argmin}_X \; -\log p(X, Y) \\ &= \mathbf{argmin}_X \; \{\, I(x_0) + \textstyle\sum_i D_{x_{\pi(i)}, x_i} \,\} \end{aligned}$$

式 (1) から，最後の式は $\mathbf{argmin}_X \; Cost(\ell, I, D, \mathcal{T})$ に等しい．よって，定理が示された． □

[系 4] 式 (2) の確率モデル $p(Z)$ において，節 3. の最適化アルゴリズム DPAO を用いて，与えられた葉の割り当て Y から，事後確率 $p(X \mid Y)$ を最大化する割り当て X を多項式時間で計算可能である．
(証明) 定理 2 と上記の定理 3 より，系が示される． □

5. 実　　験

本節では，インフルエンザウイルスの進化系統樹に，DPAO アルゴリズムを適用し，ウイルスの地理的な移動の特徴 [7] を解析した．

表 1　提案アルゴリズムにより推定された，インフルエンザウイルスの地理的移動．行は枝の親の地理ラベルを表し，列は枝の子の地理ラベルを表す．

From \ To	E-SE-Asia	Europe	N-America	Oceania	C-Asia	S-America	Africa	Middle-East	TOTAL
E-SE-Asia	2187	68	118	68	13	26	10	2	2492
Europe	60	1115	82	33	6	23	5	0	1324
N-America	70	102	2144	44	3	39	14	14	2430
Oceania	36	36	35	762	1	11	0	1	882
C-Asia	1	2	0	1	33	0	1	0	38
S-America	13	16	31	8	0	295	1	0	364
Africa	2	5	4	0	0	2	27	0	40
Middle-East	1	0	1	0	1	0	0	5	8
TOTAL	2370	1344	2415	916	57	396	58	22	7578

5.1 データ

NCBI Influenza Virus Resource から，H3N2 亜型インフルエンザウイルスの HA タンパクの遺伝子配列 3791 本を取得し，近隣結合法を用いて進化系統樹を作成した．各遺伝子配列には，分離国の情報が付加されている．分離国情報に基づき，遺伝子配列に {E-SE-Asia, Europe, N-America, Oceania, C-Asia, S-America, Africa, Middle-East} の 8 つの地理ラベルを割り当て，アルゴリズムの入力とした．

5.2 方法と結果

進化系統樹の葉 (遺伝子配列) の地理ラベルに基づき，DPAO によって最適ラベル割り当てを行い，内部頂点 (先祖配列) の地理ラベルを推定した．コスト関数には，例 1 で示した関数を用いた．結果の系統樹における枝の両端の頂点の地理ラベルを集計した分割表を表 1 に示す．

分割表はウイルスの伝播における地理的な移動を示す．提案アルゴリズムは地理ラベルの差異が少なくなるようにラベルを割り当てるので，表の対角線部分の要素数が多い．つまり，同一地域内でのウイルスの伝播が多くなっている．

次に，[4] にしたがい，対数線形モデルを用いてウイルスの地理的な移動のパラメータを推定し，モデルの検定を行う．対数線形モデルは，分割表の i 行 j 列目の期待度数 $\hat{f_{ij}}$ を

$$ln(\hat{f_{ij}}) = \theta + \lambda_i^{From} + \lambda_j^{To} + \lambda_{ij}^{FromTo} \qquad (3)$$

で表す統計モデルである．

(3) 式の λ_{ij}^{FromTo} の項を 0 とするモデルを**独立モデル**とよぶ．$i = j$ のとき $\lambda_i^{From} = \lambda_j^{To}$, $i \neq j$ のとき $\lambda_{ij}^{FromTo} = \lambda_{ji}^{FromTo}$ とするモデルを**対称モデル**とよぶ．対称モデルにおける $i = j$ のときの制約を，**周辺等分散性制約**とよぶ．対称モデルから周辺等分散性制約をなくしたモデルを**準対称モデル**とよぶ．

表 1 の分割表から，最尤推定法を用いて独立モデル，対称モデル，準対称モデルの各パラメータを推定し，χ^2 検定によりモデルの検定を行った．モデルの最尤推定および χ^2 検定には LEM [8] プログラムを用いた．

表 2 に示すように，三つのモデルとも χ^2 値と L^2 値が大きく，モデルが棄却された．どのモデルでも，中東地域の推定値が実測値から大きく外れていた．中東地域のウイルス株は，米国の兵士から分離されたものが大部分を占めているため，北米から中東への株の移動が極端に多くなっていると考えられた．そこで，中東地域を除いた分割表を用いて同様の解析を行った．表 3 に結果を示すように，中東地域を除くと分割表は準対称モデルでよくフィッティングできることが明らかとなった．

周辺等分散性制約の有無が対称モデルと準対称モデルの唯一の違いであることから，周辺等分散性がインフルエンザウイルスの地理的な移動で成り立たないことがわかる．以上より，ウイルスの移動元と移動先の地理的な分布には，有意に差があることが明らかとなった．

表 2　各対数線形モデルにおける χ^2 検定の結果

対数線形モデル	χ^2 値	L^2 値	自由度	p 値	結果
独立モデル	27339.5	14327.7	49	0.00	棄却
対称モデル	274.1	110.8	28	0.00	棄却
準対称モデル	87.1	45.6	21	0.00	棄却

表 3　中東地域を除いた場合の χ^2 検定の結果

対数線形モデル	χ^2 値	L^2 値	自由度	p 値	結果
独立モデル	26249.4	14249.9	36	0.00	棄却
対称モデル	60.8	65.9	21	0.00	棄却
準対称モデル	14.8	16.3	15	0.47	棄却しない

6. おわりに

本稿では，コスト関数が与えられた場合に，葉へのラベル割り当てから，動的計画法に基づいて木への最適ラベル割り当てを計算する多項式時間アルゴリズム DPAO を与えた．

今後の課題として，系統樹の枝の重みを考慮したアルゴリズムや，葉へのラベル割り当てだけから，コスト関数と最適ラベル割り当ての両方を求めるアルゴリズムの研究が挙げられる．

文　献

[1] A. V. Aho and J. E. Hopcroft. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, 1974.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2006.

[3] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[4] U. Engel and J. Reinecke. *Analysis of Change: Advanced Techniques in Panel Data Analysis.* Walter de Gruyter, 1996.

[5] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., 1979.

[6] W. Li and D. Graur. Fundamentals of Molecular Evolution. *Sunderland, Massachusetts: Sinauer*, 1991.

[7] C. Russell, T. Jones, I. Barr, N. Cox, R. Garten, V. Gregory, I. Gust, A. Hampson, A. Hay, A. Hurt, et al. The Global Circulation of Seasonal Influenza A (H3N2) Viruses. *Science*, 320(5874):340, 2008.

[8] J. Vermunt. LEM: A general program for the analysis of categorical data. *Tilburg University*, 1997.