Visualization and Prediction of Disease Interactions with Continuous-Time Hidden Markov Models

Jose M. Leiva-Murillo, *Member, IEEE,* and Antonio Artés-Rodríguez, *Senior Member, IEEE,* and Enrique Baca-García

Abstract—This paper describes a method for discovering disease relationships and predicting the evolution of diseases from medical records. The method makes use of continuous-time Markov chain models that overcome some drawbacks of the more widely used discrete-time chain models. The model addresses uncertainty in the diagnoses, possible diagnosis errors and the existence of multiple alternative diagnoses in the records. A set of experiments, performed on a dataset of psychiatric medical records, shows the capability of the model to visualize maps of comorbidity and causal interactions among diseases as well as to perform predictions of their future evolution.

Index Terms—Data Mining, Bioinformatics, Markov Processes.

1 Introduction

The need for large-scale statistical models have a double value in biomedicine. On one hand, they help unfold causal relationships among genetic and environmental factors in the development of diseases. Hence, biomedical research becomes increasingly dependent on data mining on large biological and medical datasets. On the other hand, the conclusions extracted from those models have a potential positive impact from a social point of view, since a more sophisticated identification of the risk factors that lead to serious disorders is made possible.

Most of the effort in biomedical data mining is focused on the discovery of genotype-phenotype interactions, which is easier when only genetic disorders are considered [5]. The huge and increasing amount of both data and scientific literature has made necessary initiatives as the Genome Ontology [3], the Disease Ontology [11], and platforms for automatic text-mining for the extraction of relevant information about genes, proteins, drugs or diseases, as well as their interactions [8], [2], [4].

Very often, much phenotypic information is available, while its underlying genetic or environmental causes remain unrecorded. In those cases, valuable information can still be extracted about the interaction of phenotypes. In the case of medical records, a causal relationship between two diseases can be uncovered if both diseases are present in the same patient at different moments along his life, and the same co-occurrence takes place in a number of patients. The best attempt to find such interactions makes use of static information, where statistical measurements about the convergence of diseases in the same patients are used to build a graph of disease relationships [6]. Statistical techniques able to model sequences of events are necessary to incorporate the time dimension to the comorbidity studies, so that the strength of the interaction between two diseases not only depend on the frequency of their co-ocurrence in the same patient, but also on the time lag between those occurrences. However, a complete characterization of a patient's clinical history as a stochastic process is intractable since it is given by the joint distribution on all the instants considered.

Consequently, the most extended tool to deal with sequential data is the discrete-time Markov chain (DTMC). In a Markov chain, a future state is statistically independent of the past, given the present. If each observation is assumed to statistically depend only on an underlying hidden state, the model is called a *hidden Markov model* (HMM). HMMs allow us to work with sequences in which, although the Markov property does not hold for observations, they become conditionally independent given the states. Because of this, they have been successfully used in speech recognition and computational biology [12].

We are interested in a more general case where the time variable is continuous. Hence, we use a continuous-time Markov chain (CTMC) model instead of a DTMC. The difficulties of estimating the parameters of a CTMC have made them appropriate in a limited set of applications, where either the number of states involved is low, or the amount of allowed transitions are far lower than the set of all possible

- *Dr. Leiva-Murillo and Prof. Artés-Rodríguez are with the Signal Processing Department, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés. E-mail: {jose, antonio}@tsc.uc3m.es*
- *Dr. Baca-García is with the Fundación Jiménez-Díaz, Avda. Reyes Católicos 2, 28040 Madrid. E-mail eb2452@columbia.edu*

transitions. An example of the former is the model with 4 states used in evolutionary biology to model sequence of nucleotides [14], and a reparameterization is proposed to further reduce the number of parameters involved. In systems biology, the number of states is denumerable when modeling the size of a species population, but the state describing a population of N individuals can only jump to the $N-1$ and the $N+1$ states [7].

In this paper, we propose a novel maximum-likelihood approach to the estimation of CTMC parameters by treating the model as a probability density function on t. This way, no limitation on the number of states or transitions exists, as far as the availability of enough data guarantees that no risk of overfitting exist. In order to address the uncertainty in the diagnoses [9], we study the embedded HMM in which observations are assumed to depend on underlying hidden states in a probabilistic way [12]. This way, the CTMC is extended to a continuous-time HMM (CT-HMM) that incorporates hidden states (the ground-truth diseases) that are de-coupled from the observations (the diagnoses given by doctors). The *emission* parameters of the model describe the probability of a given disease to be diagnosed in a given way, while the *transition* parameters describe how often and intensively in time a transition between two states occur. This transition parameters can give us a measure of either comorbid or causal relations among diseases.

The model presented in the paper makes use of the following assumptions:

- The different patients are instances of the same stochastic process to be modeled.
- The process is stationary, so that the intensity of the interaction between two diseases does not depend on the age of the subject.
- A patient stays in the same state (disease) until the time instant of the following medical claim.
- The Markov property holds, and this implies that the future clinical history of a patient only depends on his/her present state, and is independent of the past.

The paper is organized as follows. In Section 2, the basic theory about CTMCs is provided, together with a maximum-likelihood procedure to obtain the parameters of the model. In Section 3 the model is extended to account for hidden states, and describe how different elements from HMMs and CTMCs can be dealt with in a unified framework. In Section 4, a model is trained with a database of medical records from psychiatric patients, and we describe how the model can be used to build a map of disease interactions, as well as to predict the future evolution of a patient, given the current disorder. Finally, in Section 5 we describe the most relevant conclusions of this work and propose a set of potential extensions and applications.

2 Continuous Time Markov Chains

Let $\mathcal{S} = \{1, 2, \ldots, N_S\}$ be a finite set of states. Let us assume that those states are fully observable. We consider a stochastic process that

- stays in state i for a time distributed according to $F_i(t)$.
- moves to state j with probability p_{ij} at the end of this holding time.

If the random process fulfills the Markov property, that is: $P(X(s+t)|X(u)) = P(X(s+t)|X(s))$ with $u \leq s$, then the process is a continuous-time Markov chain (CTMC).

In the following, we describe how the CTMC can be characterized as a probability density function on t, which allows us to find the parameters of the model by maximum-likelihood.

2.1 Characterization of CTMC as a probability density function on t

In a continuous time Markov chain (CTMC), [7], [13], the holding time is exponentially distributed according to $F_i(t) = e^{-q_i t}$, where q_i is the *intensity* of the transitions out of state i, i.e. the tendency of the process to leave state i. Let us define the *embedded* Markov chain as the discrete-time process resulting of taking the sequence of states $S = \{s_n\}, n = 0, \ldots, L$ at the state transitions. With the support of the embedded chain, and following the notation in Ref. 7, we get

$$P[s_n = j, \tau_n > t | s_{n-1} = i] = p_{ij} e^{-q_i t}. \tag{1}$$

where τ_n is the time lag from the $n-1$ to the n-th transition; and $p_{ij} = q_{ij}/q_i$, being q_{ij} each of the elements of the so-called infinitesimal generator matrix Q. Each q_{ij} describes the intensity of the transition from i to j, with $\sum_j q_{ij} = q_i$ [7], Th. 4.7.

Note that the probabilistic model (1) is a probability mass function for the discrete component s_n, and a cumulative distribution function for t. As a consequence, we can define a probability density function on t, denoted as $f_{ij}(t)$ for $i \to j$ transitions. Its value is given by

$$\begin{aligned} f_{ij}(t) &= f(\tau_n = t, s_n = j | s_{n-1} = i, Q) \\ &= \frac{d}{dt} P[s_n = j, \tau_n > t | s_{n-1} = i] = q_{ij} e^{-q_i t}. \end{aligned} \tag{2}$$

Then, the model is completely defined by the matrix $\{q_{ij}\}$. We now describe the procedure for obtaining the values of q_{ij} that best fit to data.

2.2 Maximum-Likelihood optimization for Q

Let $T = \{\tau_n\}$ be the sequence of observed time lags between successive transitions. Thanks to the Markov

property, the joint density for all those transitions can be factorized as

$$f(T, S|Q) = f(\tau_0, s_0) \prod_{n=1}^{L} f(\tau_n, s_n|s_{n-1}, Q).$$

The maximization of the log-likelihood leads to

$$Q^* = \arg\max_Q \log f(T, S|Q) \tag{3}$$

$$= \arg\max_Q \left[\log f(\tau_0, s_0) + \sum_{n=1}^{L} \log f(\tau_n, s_n|s_{n-1}, Q) \right]. \tag{4}$$

We obtain each q_{ij} by finding the point at which

$$\frac{d}{dq_{ij}} \log f(T, S|Q) = 0.$$

It easy to check that the solution to this problem is given by

$$q_{ij} = \frac{N_{ij}}{\sum_{n:s_n=i} \tau_n},$$

where N_{ij} is the number of observed transitions from i to j.

The procedure described so far assumes that states are observed without uncertainty. In a medical application, this is equivalent to assume that diagnoses provided by the doctors are noise-free and specific enough. In the following we extend the model to treat observations in a probabilistic way with respect to the underlying (ground-truth) states, in order to account for diagnosis errors and lack of detail in the diagnoses.

3 Continuous-Time Hidden Markov Model

We now consider the case where the sequence of states is not fully observed. In a clinical record, for example, de-coupling observations from states can lead to a model more robust to diagnosis errors. The set of possible observations can contain the set of states, with the inclusion of undetermined or under-defined diagnoses. Another option is to consider a set of states that have no correspondence with a-priori known diseases. Assuming that the set of observations is also finite, $\mathcal{O} = \{1, 2, \ldots, N_O\}$ a matrix of parameters B must be added to the model, with $b_i(k) = P(o_n = k|s_n = i)$. This results in assuming that the embedded discrete-time Markov chain is a hidden Markov model (HMM).

3.1 Optimization of the CT-HMM parameters

The parameters of the model are given by the set $\lambda = (\{\pi_i\}, \{q_{ij}\}, \{b_i(k)\})$, and the joint probability of a sequence of observations O and a sequence of states S is given by the following factorization, given that s_n only depends on s_{n-1} and o_n only depends on s_n

$$\begin{aligned} P(O, S|\lambda) &= \pi_{s_0} \prod_n f(\tau_n, s_n|s_{n-1}) b_{s_n}(o_n) \\ &= \pi_{s_0} \prod_n a_{s_{n-1}s_n}(\tau_n) b_{s_n}(o_n), \end{aligned} \tag{5}$$

where we have used the short-hand notation $a_{s_{n-1}s_n}(\tau_n) = q_{s_{n-1}s_n} e^{-q_{s_{n-1}}\tau_n}$ to make the connection with discrete-time HMM's literature more clear. The parameters can be adjusted from the observations by means of the Expectation-Maximization algorithm [12], [1]. The objective is to solve the following optimization problem

$$\arg\max_\lambda \mathcal{L}(\lambda, \lambda^0) = \arg\max_\lambda E_{P(O,S|\lambda^0)}[\log P(O, S|\lambda)], \tag{6}$$

being λ^0 an initial set of parameters. The iteration of (6) and the update of λ^0 lead to a local optimum. The likelihood $\mathcal{L}$ can be decomposed in terms involving the different sets of parameters, the following way

$$\begin{aligned} \mathcal{L} = &\sum_{S\in\mathcal{S}} P(O, S|\lambda^0) \log \pi_{s_0} \\ &+ \sum_{S\in\mathcal{S}} P(O, S|\lambda^0) \sum_n \log a_{s_{n-1}s_n}(\tau_n) \\ &+ \sum_{S\in\mathcal{S}} P(O, S|\lambda^0) \sum_n \log b_{s_n}(o_n). \end{aligned} \tag{7}$$

Note that for $\{\pi_i\}$ and $\{b_i(k)\}$ the updates are as in a conventional, discrete-time HMM. However, we must work out an update rule for the continuous-time component involving the $\{q_{ij}\}$. We can express the second term in (7) as

$$\mathcal{L}_Q = \sum_{i,j} \sum_n P(O, s_{n-1} = i, s_n = j|\lambda^0) \log a_{s_{n-1}s_n}(\tau_n). \tag{8}$$

Reaching the fixed-point described by $\partial\mathcal{L}(Q)/\partial q_{ij} = 0$, leads to the following update rule for q_{ij}

$$\begin{aligned} q_{ij} &= \frac{\sum_{n=1}^{L} P(O, s_{n-1} = i, s_n = j|\lambda^0)}{\sum_{n=1}^{L} \tau_n P(O, s_{n-1} = i|\lambda^0)} \\ &= \frac{\sum_{n=1}^{L} \alpha_{n-1}(i) a_{ij}(\tau_n) \beta_n(j) b_j(o_n^{(m)})}{\sum_{n=1}^{L} \tau_n \alpha_{n-1}(i) \beta_n(i)}, \end{aligned} \tag{9}$$

where $\alpha_n(i) = P(o_0 o_1 \ldots o_n|s_n = i)$ and $\beta_n(i) = P(o_{n+1} o_{n+2} \ldots o_L|s_n = i)$ are the forward and backward auxiliary variables, respectively [12]. Here we have not included the update rules for $\{\pi_i\}$ and $\{b_i(k)\}$ because they are identical to the ones given for discrete-time HMMs, as far as the sequences $\{\alpha_n(i)\}$ and $\{\beta_n(i)\}$ are now computed in continuous time.

3.2 Working with multiple sequences

Very often, a set of sequences $\{S_m\}, m = 1, \ldots, N_m$, each with length L_m, is available for training the

model instead of a single one. In that case, the rule (9) becomes

$$q_{ij} = \frac{\sum_{m=1}^{N_m} \sum_{n=1}^{L_m} \hat{\alpha}_{n-1}^{(m)}(i) a_{ij}(\tau_n^{(m)}) \hat{\beta}_n^{(m)}(j) b_j(o_n^{(m)})}{\sum_{m=1}^{N_m} \sum_{n=1}^{L_m} \tau_n^{(m)} \hat{\alpha}_{n-1}^{(m)}(i) \hat{\beta}_n^{(m)}(i)}, \tag{10}$$

where $\hat{\alpha}_n^{(m)}(i)$ and $\hat{\beta}_n^{(m)}(i)$ are the normalized forward and backward variables such that $\sum_i \hat{\alpha}_n^{(m)}(i) = 1$ and $\sum_i \hat{\beta}_n^{(m)}(i) = 1$ (see [12], Sec. V-A).

3.3 Working with simultaneous observations

We are interested in a more general model that allows more than just one observation to take place at the same time. In terms of the application at hand, the motivation comes from the fact that a doctor can give several alternative diagnoses to a given patient, according to a given set of symptoms. We deal with this situation in a probabilistic way: when n several alternative diagnoses are provided, we assume that each of them is actually observed with probability $1/n$. This way, both primary and secondary (when present) diagnosis can be considered.

3.4 Prediction with CTMCs

So far, we have worked with a CTMC model in which $p_{ii} \geq 0$ and $q_{ii} \geq 0$ which allows a process, in a transition, to *jump* to the same state. However, in the standard CTMC framework, $p_{ii} = 0$ and $q_{ii} = -\sum_{k \neq i} q_{ik}$. In other words, in a CTMC, transitions from one state to the same one do not make sense. We can obtain an equivalent CTMC in which the parameters are standardized. The new intensity for transitions from S_i is given by $q_i' = \sum_{k \neq i} q_{ij} = q_i - q_{ii}$, i.e. the intensity is lower because in the new (standard) formulation there are fewer transitions - the transitions to the same state are *removed*. The equivalence between both models are given by the equations

$$\begin{aligned} q_{ij}' &= q_{ij} & p_{ij}' &= \frac{p_{ij}}{1 - p_{ii}}, \qquad i \neq j \\ q_{ii}' &= -\sum_{k \neq i} q_{ij} & p_{ii}' &= 0. \end{aligned}$$

Let $P(t)$ be the matrix with components $P_{ij}(t) = P(s_n = j | s_{n-1} = i, \tau_n = t)$, i.e. it gives us the probability of being at state j in a future instant t given that the current state is i. The new structure of matrix Q allows for expressing $P(t)$ in terms of the Euler matrix exponential

$$P(t) = \exp(Qt). \tag{11}$$

Let the initial distribution of probabilities across states be given by the vector $\pi(0) = \{\pi_i(0) = P(X(0) = i)\}$, and $\pi(t)$ be the distribution vector in instant t with components $\{\pi_i(t) = P(X(t) = i)\}$. Then, the following equality gives us the ability of performing prediction

$$\pi(t) = P(t)^T \pi(0). \tag{12}$$

4 Results

In the following, we present a set of experiments performed on a dataset with 6,995,364 clinical visits. This dataset has been recorded from $374,955$ patients for over 30 years in different locations of the Community of Madrid, Spain.

We use the categorization of the World Health Organization (WHO) of health disorders under the International Classification of Diseases (ICD10) [10]. The chapter F is devoted to mental and behavioral disorders. We truncate the codes at the second layer, so that digits beyond the third one are not considered. As a result, we build a model with 74 states (68 of them corresponding to mental disorders) and 136 possible observations (in addition to the states, we considered the null diagnosis, the under-defined diagnoses F1, F3, F4 and F5, and medical factors from the Z family).

In the following we describe the two main applications of the model: the visualization of a map of disease interactions and the prediction of the future evolution of a patient, given his current disorder.

4.1 Visualization of the model

According to our model, there are two ways to measure the interaction intensity between disorders. The first one is through the study of matrix $Q = \{q_{ij}\}$. The resulting graph takes into account the time dimension, i.e. q_{ij} gives us information about how fast the transition between states i and j is, if the transition takes place.

The second option is to represent the elements of matrix P, whose elements have been normalized as $\{p_{ij} = q_{ij}/q_i\}$. These values can be interpreted as the transition coefficients of an embedded HMM, i.e. the Markov model which ignores the length of time periods between medical claims.

We show the graph obtained from both sets of coefficients $\{q_{ij}\}$ and $\{p_{ij}\}$ in Figs. 1 and 2, respectively. In both cases, the highest values of each matrix have been extracted. The thickness and color of arrows express the intensity of the interaction between each pair of diseases. The size of the nodes is proportional to the presence of each disease in the dataset.

The exploration of the graph leads to conclusions that are consistent with the medical knowledge. In both graphs, mood disorders (F34), anxiety (F41) and stress disorders (F43) appear as *absorbing* states - they have multiple causes but rarely lead to other disorders. Also, the presence of other non-mental diseases such as blindness (H54), epilepsy (G40) and virus infections of the nervous system (A81) reveals causal relationships with mental disorders that deserves further medical analysis.

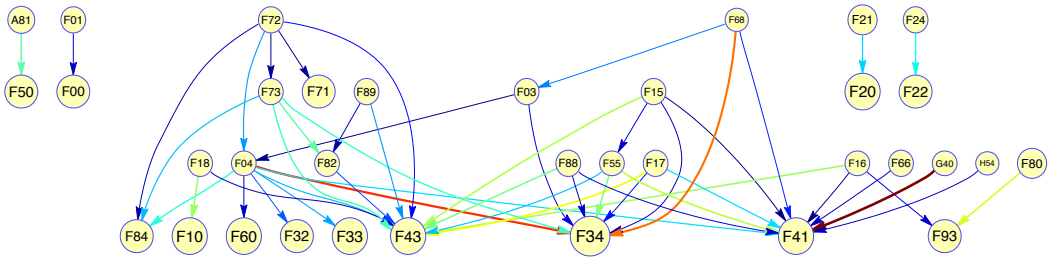


Fig. 1. Representation of the strongest interactions given by matrix $\{q_{ij}\}$.

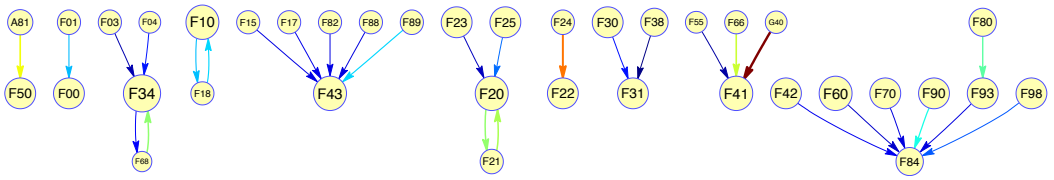


Fig. 2. Representation of the strongest interactions given by matrix $\{p_{ij}\}$.

We can conclude that matrix P gives us the most valuable information about how closely related the different pairs of diseases are. However, once a causal relationship is discovered, its corresponding link in matrix Q must be explored to know how fast the transition takes place along time. However, we must stress that only when $P(t)$ is computed according to (11), the complete probabilistic characterization along time is obtained.

4.2 Prediction of future evolution of disorders

We have described in Sec. 3.4 how matrix Q contains all the necessary information to predict the state (disorder) at a future instant given the current disease. We show in Figs. 3 to 5 three different cases in which, depending on the current disease, the model predicts that either i) the disease will stay stable; ii) the disease will evolve to another disease with high probability, or iii) the disease is likely to evolve to a set of alternative diseases.

We show the estimated distribution of probabilities for the different diseases in the time period ranging from one month to four years, obtained according to (12) and an a-priori distribution concentrated in the indicated initial disease.

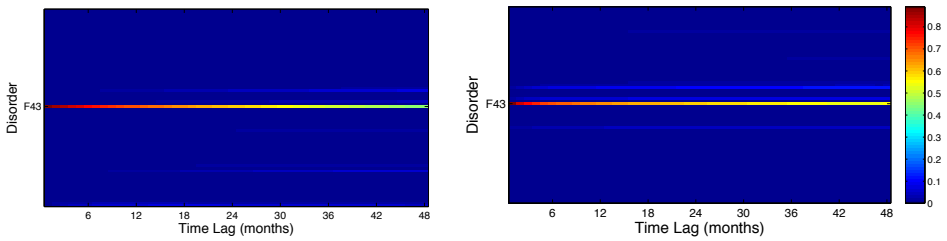
To validate the prediction ability of the model, the probabilities provided by (12) are compared with the empirical probabilities obtained from the frequencies directly observed in the database. To do so, the patients that have been diagnosed with the disease of interest are collected, and their real future 4-years trajectories are drawn in the states-vs-time matrix. The resulting matrix is normalized according to $\sum_j P_{ij}(t) = 1$.

The results are shown in Figures 3 to 5. In Fig. 3, the evolution of F43 (reaction to severe stress) is studied. Fig. 3(a) predicts a stability along time. This prediction shows a high degree of agreement with the real situations displayed in Fig. 3(b). Fig. 4 analyzes a case in which a disorder A81 (virus infections of the central nervous system) evolves with high probability to another disorder: F50 (eating disorders). Again, the model predicts the real trajectories of patients with high accuracy. Finally, in Fig. 5, F15 (mental disorders due to use of stimulants) is considered. In Fig. 5(a), the model predicts that it can be a causal factor in a number of disorders, including F34 (persistent mood disorders), F41 (anxiety), F43, and F60 (personality disorders). Those transitions do actually take place, as confirmed by Fig. 5(b).

We can observe that the predictions are in agreement with the graphs in Figs. 1 and 2. The stability shown by disorder F43 is due to its *absorbing* nature -no transitions take place from F43 in none of the graphs. The transition from A81 to F50 appears explicitly in both graphs. Finally, the model predicts a high probability of transition from F15 to F34, F41 and F43 which is both suggested by the graph and validated by the observations. However, the predicted transitions from F15 to F84 and F93 are due to second or higher order interactions in the graph and are not supported by the observations.

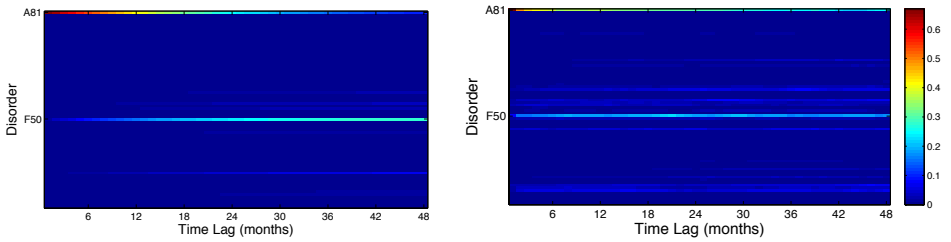
5 Conclusions

We have described a CT-HMM model able to discover not only the relationships among diseases, but also, and more importantly, the intensity of those interactions along time. We have shown how the model can be represented as a map of disease interactions whose interpretation can lead to relevant medical interpretations. Because the model allows for the prediction of future transitions to other diseases from a present one, its use as a tool for the generation of risk alarms is of



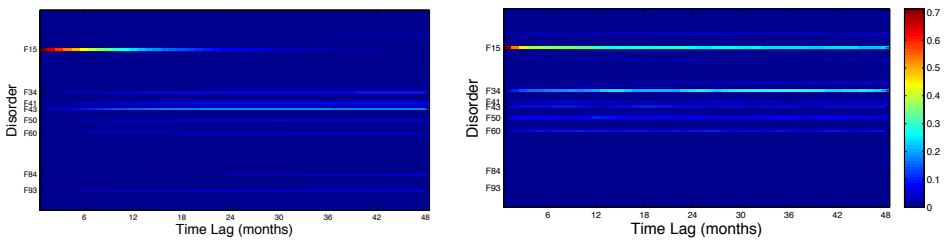
(a) Prediction for F43. (b) Real trajectories for F43.

Fig. 3. Evolution predicted (left) and observed (right) for a stable disorder.



(a) Prediction for A81. (b) Real trajectories for A81.

Fig. 4. Evolution predicted (left) and observed (right) for a disorder with a strong causal link to another one.



(a) Prediction for F15. (b) Real trajectories for F15.

Fig. 5. Evolution predicted (left) and observed (right) for a disorder that can evolve to a number of other disorders.

potential interest from the point of view of medical care.

Further medical conclusions can be extracted if we do not impose a correspondence between the set of states considered and the ones present in standard taxonomies. In this case, the resulting model can help validate a given taxonomy, by analyzing which observed disorders are linked to the same underlying state.

The method described can be the first building block from which clustering and classification tasks can be performed on patients, modeled by their corresponding sequences of states.

REFERENCES

[1] Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report tr-97-021, International Computer Science Institute, Berkeley, California, USA, 1997.

[2] Jung-Hsien Chiang and Hsu-Chun Yu. Literature extraction of protein functions using sentence pattern mining. *IEEE Trans. Knowl. Data Eng*, 17(8):1088–1098, 2005.

[3] Harris M. Clark et al. The gene ontology database and informatics resource. *Nuclei Acid Research*, 32:D258–261, 2004.

[4] Oana Frunza, Diana Inkpen, and Thomas Tran. A machine learning approach for identifying disease-treatment relations in short texts. *IEEE Trans. Knowl. Data Eng*, 23(6):801–814, 2011.

[5] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi. The human disease network. *PNAS*, 104(21):8685–8690, 2007.

[6] Cesar A. Hidalgo, Nicholas Blumm, Albert-Laszlo Barabasi, and Nicholas Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4), 2009.

[7] Massaki Kijima. *Markov Processes for Stochastic Modeling*. Chapman and Hall, Cambridge, UK, 1997.

[8] Martin Krallinger, Florian Leitner, and Alfonso Valencia. Analysis of biological processes and diseases using text mining approaches. *Methods in Molecular Biology*, 593:341–382, 2010.

[9] Daniel Nikovski. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Trans. Knowl. Data Eng*, 12(4):509–516, 2000.

[10] World Health Organization. *Family of International Classifications*. 2011.

[11] John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin2and Warren A Kibbe, Lihua Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm. Annotating the human genome with disease ontology. *BMC Genomics*, 10(Suppl. 1), 2010.

[12] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[13] Sheldon M. Ross. *Stochastic Processes, 2nd Edition*. Wiley, 1996.

[14] Marc A. Suchard, Robert E. Weiss, and Janet S. Sinsheimer. Bayesian selection of continuous-time markov chain evolutionary models. *Molecular Biology and Evolution*, 32(6):1001–1013, 2001.