This is an Accepted Manuscript version of the following article, accepted for publication in **CYBERNETICS AND SYSTEMS**.
Postprint of: Wang M., Lai Y., Li M., Zhang H., Szczerbicki E., Toward Human Chromosome Knowledge Engine, CYBERNETICS AND SYSTEMS (2023), pp. 1-8, DOI: 10.1080/01969722.2022.2162743

Towards Human Chromosome Knowledge Engine

Maiqi Wang [a], Yi Lai [b], Minghui Li [a], Haoxi Zhang [a], Edward Szczerbicki [c]

[a] School of Cybersecurity, Chengdu University of Information Technology, Chengdu, China; [b] Department of Medical Genetics/Prenatal Diagnostic Center, West China Second Hospital, Sichuan University, Chengdu, China; [c] Faculty of Management and Economics, Gdansk University of Technology, Gdansk, Poland

Abstract. Human chromosomes carry genetic information about our life. Chromosome classification is crucial for karyotype analysis. Existing chromosome classification methods do not take into account reasoning, such as: analyzing the relationship between variables, modelling uncertainty, and performing causal reasoning. In this paper, we introduce a knowledge engine for reasoning-based human chromosome classification that stores knowledge of chromosomes via a novel representation structure, the Chromosome Part Description (CPD), and reasons over CPDs by utilizing the probability tree model (PTM) for classification. Each CPD keeps information on a particular feature of chromosomes, while the PTM provides causal reasoning capability taking CPDs as nodes and dependencies between CPDs and types as edges. Experimental results show that the proposed knowledge engine's performance increases when providing more CPDs and achieves 100% classification accuracy with more than three CPDs.

Address correspondence to Haoxi Zhang, School of Cybersecurity, Chengdu University of Information Technology, No. 24 Block 1, Xuefu Road, Chengdu, China, 610225. E-mail: haoxi@cuit.edu.cn

Keywords: Knowledge engine, human chromosomes, probability tree model, Chromosome Part Description, causal reasoning.

INTRODUCTION

Chromosome classification plays a critical role in karyotype analysis. Human chromosomes are the carriers of human genetic materials and genes, and karyotype analysis is an important technique to identify genetic abnormalities through chromosome metaphase images. Karyotype analysis is carried out by preparing karyotype images through segmenting metaphase images and then classifying and organizing chromosome instances into 23 pairs, including 22 pairs of autosomes and a pair of sex chromosomes (XY for males and XX for females) and sending the prepared karyotype images to experts for final analysis (Piper&Granum, 1989).

The study of chromosome classification appeared as early as the end of the last century. Jenq et al. (1992) proposed a method of central axis transformation as a pre-process to help the classification performance. Lerner et al. (1995) try to use neural networks to complete the task of chromosome classification. Ritter et al. (1997) use chromosome length and centromere position information to classify chromosomes. However, due to the complex characteristics of chromosomes and the limitations of the technology at that time, these early methods were highly dependent on geometrical features (e.g., the chromosome's axis, length, and centromere position) and could hardly achieve satisfactory results in accuracy. Kusakci et al. (2017) proposed a method for chromosome classification based on multiple support vector machines, which followed the same technique path as previous studies. Oskouei et al. (2010) proposed a chromosome classification method based on a wavelet neural network, which uses chromosome size and the proportional density distribution of long and short arms as

feature vectors and achieves good accuracy.

However, the methods mentioned above do not take into account reasoning, such as: analyzing the relationship between variables, modelling uncertainty, and performing causal reasoning. Hence, they are barely explainable in their classification outcomes. In real clinic scenarios, cytogeneticists need to know how and why a chromosome image is classified, which requires reasoning capability and interpretability that those methods do not provide. In order to solve these problems, we propose the knowledge engine for human chromosome classification and introduce a novel representation structure, Chromosome Part Description (CPD), to retain the various features of chromosomes and utilize the probability tree model (PTM) to represent causality and support causal reasoning. The occurrence of related CPDs serves as nodes in the PTM, and the edges in the PTM are causal relationships between features and types. Finally, causal reasoning methods (Pearl, 2000), such as conditional probability, intervention, and counterfactual reasoning, are utilized to carry out the identifying results of a given chromosome image.

KNOWLEDGE AND PROBABILITY TREE MODEL

Knowledge is a familiarity, awareness, or understanding of someone or something, such as facts (propositional knowledge), skills (procedural knowledge), or objects (acquaintance knowledge) (Boghossian, 2007). By most accounts, knowledge can be acquired in many different ways and from many sources, including but not limited to perception, reason, memory, testimony, scientific inquiry, education, and practice (Steup, 2007). In this work, we refer knowledge to cytogeneticists' expertise in identifying chromosomes based on the features and features' relationships, and we use the CPD and PTM to do the knowledge representation and reasoning.

The probability tree, as known as the staged tree model (Görgen, 2017), is one of the fundamental models for representing the causal generative process of a random experiment or stochastic process (Genewein et al., 2020) (see Figure 1).

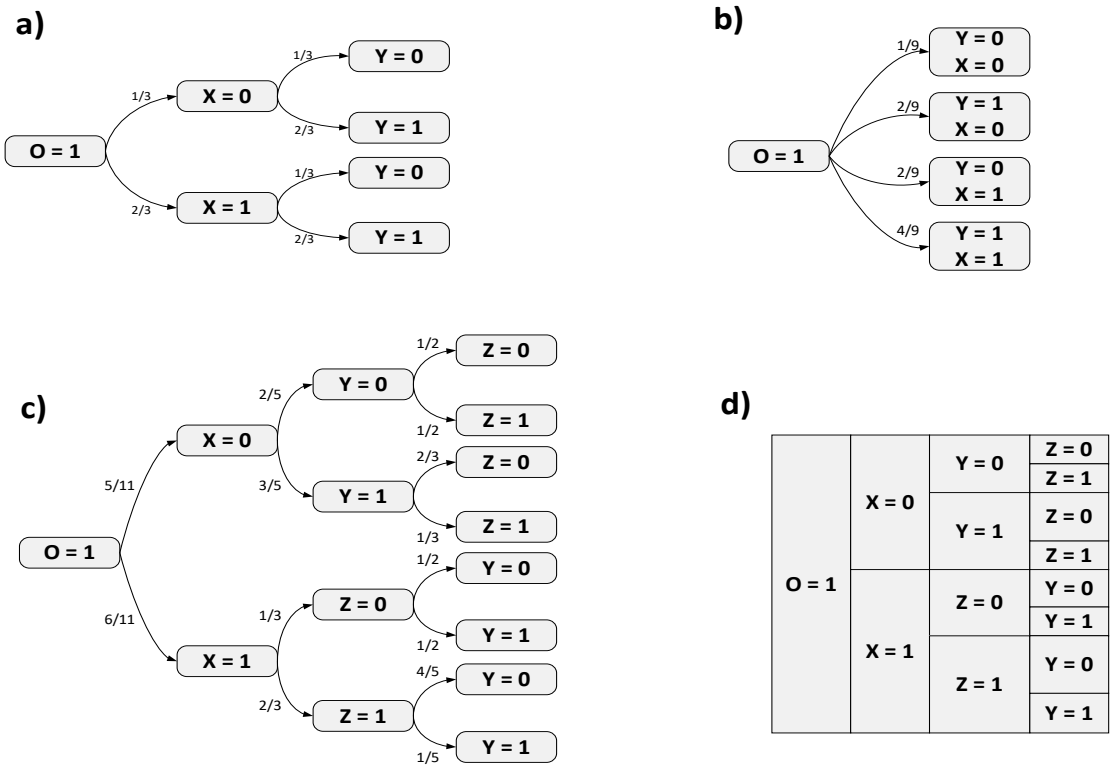

Figure 1: Probability trees.

As shown in Figure 1, panels (a) and (b) show the same joint distribution over X and Y. They differ in that (a) assumes $X \rightarrow Y$, whereas (b) does not assume a causal dependency. (c) is a more complex example of a probability tree model. It is a probability tree where $Y \rightarrow Z$ when $X = 0$ and $Z \rightarrow Y$ when $X = 1$. Panel (d) shows a probability tree mass diagram, an alternative representation of the probability tree. By convention, we bind $O = 1$ (O stands for omega "Ω" representing the sample space) at the root node.

For knowledge with causal dependences, the machine learning methods or neural networks can hardly represent and reason over them. Instead, the PTM naturally contains causal dependences (Dasgupta et al., 2019) and works with various reasoning

algorithms, such as joint probability, conditional probability, intervention operation, and counterfactual reasoning. PTM possesses clean semantics and can represent context-specific causal dependencies, which are crucial for causal induction (Genewein et al., 2020). The semantics are self-explanatory: each node in the tree corresponds to a potential state of the process, and the arrows indicate both the probabilistic transitions and the causal dependencies between them.

HUMAN CHROMOSOME KNOWLEDGE ENGINE

A. Chromosome Part Description (CPD)

This section introduces the CPD, a novel chromosome representation method based on cytogenetics expertise. It contains the characteristics of a chromosome by describing the features of the chromosome's three parts: centromere, *q*-arm, and *p*-arm.

The centromere is a specific feature on chromosomes. Chromosomes can be divided into three groups based on the position of the centromere: metacentric chromosome, submetacentric chromosome, and acrocentric chromosome (Moradi, 2003). In addition, the position of centromeres splits a chromosome into two parts, the longer part is called the *q*-arm, and the shorter part is the *p*-arm (Hanamura, 2021).

A CPD carries a feature description of one chromosome part. In our proposed method, a description consists of two entities and a relation (entity_1, relation, entity_2). There are three relations (HAS, IS, LOCATED IN) and two types of entities (part entity and feature entity). Part entities include *p*-arm, *q*-arm and centromere, while feature entities consist of deep-band, shallow-band, variation, constriction, and centromeric position. For example, we can define the CPD as (*p*-arm, has, variation) to describe a chromosome's *p*-arm has variation. Moreover, a chromosome part may have more than one feature. In other words, each part can have multiple CPDs.

B. Knowledge Base

According to cytogenetics' knowledge about chromosome features and types, we build the PTMs for each type and integrate the PTMs as a knowledge base. Here, chromosome features are defined as CPDs and taken as binary nodes (0 or 1) or observational variables in the PTMs representing the occurrence of the features, while edges are probabilities between features and types calculated based on clinic data. Let Y be the outcome variable, and $(X_1, X_2, \dots, X_n)$ represent the CPDs of a chromosome. The probability of a given chromosome is Y can be expressed as $P(Y|X_1, X_2, \dots, X_n)$ (see Figure 2).

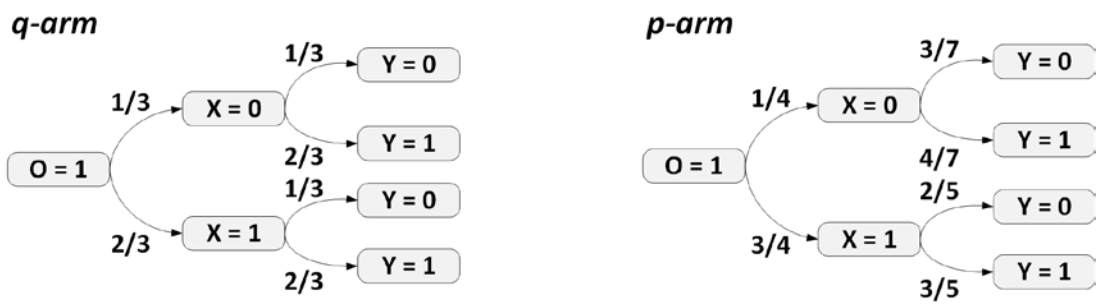

Figure 2: The probability tree models constructed with CPDs. The leaf nodes represent the final reasoning outcomes, such as *Chromosome Type*, and the ancestor nodes represent the occurrence of any related features. And the probability tree models of each chromosome type consists two trees for the *p*-arm and the *q*-arm respectively.

To simplify the PTMs, we build two probability trees for any given type: a *q*-arm tree and a *p*-arm tree, and the features related to centromeres are embedded into both trees. Dividing into two trees can significantly reduce the number of nodes in the tree due to the constraint of the causal reasoning algorithms (Genewein et al., 2020) on variables, which requires the number and type of variables between branches of the tree must be the same.

Furthermore, the built PTMs are validated using counterfactual reasoning (Kusner et al., 2017). In a probability tree, a counterfactual is a statement about a subjunctive

(*i.e.*, possible or imagined) event that could happen had the stochastic process taken a different course. This operation allows evaluation probabilities of the form $P(A_C \mid B)$. That is, "Given that *B* is true, what would the probability of *A* be if *C* were true?". Here, A_C denotes the subjunctive event *A* under the counterfactual assumption that event *C* has occurred (*i.e.*, a potential response), and *B* is the indicative (*i.e.*, factual) assumption. For our chromosome PTMs, we compare the conditional probability $P_c(\, Y_C \mid (\, X_1\, , X_2, \dots , X_n)\,)$ with $P(Y|X_1, X_2, \dots , X_n)$, if P_c is smaller, the probability tree is correct. Otherwise, the CPD corresponding to counterfactual assumption *C* is unnecessary and hence shall not be a node in the PTM.

C. Inference

This section introduces chromosome classification based on the PTMs with causal reasoning. The inference consists of three steps: 1) infer the given chromosome's conditional probabilities for each type, 2) score the probabilities for each type, and 3) determine the chromosome type according to the highest score (Figure 3).

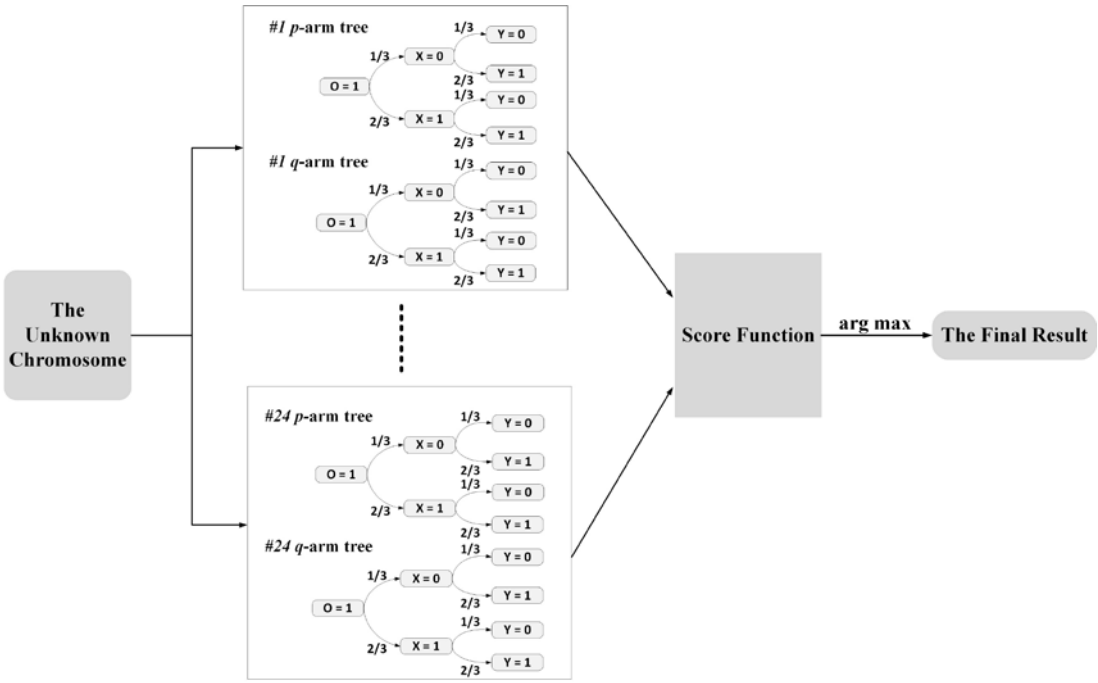

Figure 3: The main process of chromosome classification.

Specifically, given the CPDs of an unknown chromosome X_{new}, one can measure the probability of the unknown chromosome being type m by:

First, inference f_p and f_q through type-m's p-arm tree and q-arm tree via interventions (Lattimore et al., 2016), respectively, as follows:

$$X_{new} = (x'_1, x'_2, \dots, x'_n) \tag{1}$$

$$M = (1, 2, \dots, 24) \tag{2}$$

$$f_p(X_{new}, m) = P_p(Y_m | X_{new}), \qquad m \in M \tag{3}$$

$$f_q(X_{new}, m) = P_q(Y_m | X_{new}), \qquad m \in M \tag{4}$$

where M represents the set of chromosome types, a total of 24 classes.

Then, score the probability of the unknown chromosome being type m as the mean of the two inference outcomes:

$$Score(X_{new}, m) = \frac{f_p(X_{new}, m) + f_q(X_{new}, m)}{2} \tag{5}$$

Finally, the type of the unknown chromosome can be determined by inferring the probability scores for all 24 types and assigning it to the type that produces the most significant score:

$$Y = \arg\max_{m \in M} Score(X_{new}, m) \tag{6}$$

THE INITIAL EXPERIMENT

To examine the performance of our knowledge engine in identifying chromosome types, we conduct the initial experiment. For each type, we randomly pick CPDs and send them incrementally to our engine to evaluate the knowledge engine's classification results. Our primary concerns are the effects of different amounts of CPDs and certain CPDs on accuracy. Table 1 shows the results of the initial experiment.

Table 1: The results of our initial experiment.

Chromosome Type	Amount of CPDs	Accuracy
Chromosome 1	1	30%
	2	66%
	3	100%
Chromosome 2	1	43%
	2	60%
	3	100%
…	…	…
Chromosome Y	1	45%
	2	68%
	3	100%

As Table 1 shows, we learn that the accuracy improves significantly as the amount of CPDs increases. When there are three or more CPDs, the proposed knowledge engine can identify the chromosome type precisely with 100% accuracy. Since we pick the CPDs randomly, the results suggest that the engine can work robustly with different CPDs and does not depend on certain CPDs.

CONCLUSIONS AND FUTURE WORK

This paper introduces the knowledge engine for human chromosome classification. We propose a novel representation structure, Chromosome Part Description, for chromosome feature representation and utilize the probability tree model for classification. Experimental results show that the proposed knowledge engine achieves 100% classification accuracy with more than three CPDs, suggesting that our knowledge engine is promising.

As this research is at its early stage, further research and refinement remain to be done:

1) Further design and development of the CPDs: add more entities and relations and introduce n-tuples (n-greater than 3) to represent more complex features.
2) Further design and development of the probability tree model: add more causal relations and extend to first-order logic to enrich the relationship among entities and explore temporal logic (Øhrstrøm & Hasle, 2007) and probabilistic programming (Brémaud, 2012) for expanded reasoning functionality.
3) Refinement and further development of the chromosome type inference algorithms: introduce temporal information to probabilistic inference.

ACKNOWLEDGEMENT

The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions on this paper. This work was supported by the Kemoshen Science Research Program.

REFERENCES

Boghossian, P. (2007). Fear of knowledge: Against relativism and constructivism. Clarendon Press.

Brémaud, P. (2012). An introduction to probabilistic modeling. Springer Science & Business Media.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., ... & Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. arXiv preprint arXiv:1901.08162.

Genewein, T., McGrath, T., Déletang, G., Mikulik, V., Martic, M., Legg, S., & Ortega, P. A. (2020). Algorithms for causal reasoning in probability trees. arXiv preprint arXiv:2010.12237.

Görgen, C. (2017). An algebraic characterization of staged trees: their geometry and causal implications (Doctoral dissertation, University of Warwick).

Hanamura, I. (2021). Gain/amplification of chromosome arm 1q21 in multiple myeloma. Cancers, 13(2), 256.

Jenq, J. F., & Sahni, S. (1992). Serial and parallel algorithms for the medial axis transform. IEEE Transactions on Pattern Analysis & Machine Intelligence, 14(12), 1218-1224.

Kusakci, A. O., Ayvaz, B., & Karakaya, E. (2017). Towards an autonomous human chromosome classification system using Competitive Support Vector Machines Teams (CSVMT). Expert Systems with Applications, 86, 224-234.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in neural information processing systems, 30.

Lattimore, F., Lattimore, T., & Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. Advances in Neural Information Processing Systems, 29.

Lerner, B., Guterman, H., Dinstein, I., & Romem, Y. (1995). Medial axis transform-based features and a neural network for human chromosome classification. Pattern Recognition, 28(11), 1673-1683.

Moradi, M., Setarehdan, S. K., & Ghaffari, S. R. (2003, June). Automatic locating the centromere on human chromosome pictures. In 16th IEEE Symposium Computer-Based Medical Systems, 2003. Proceedings. (pp. 56-61). IEEE.

Oskouei, B. C., & Shanbehzadeh, J. (2010, December). Chromosome classification based on wavelet neural network. In 2010 International Conference on Digital Image Computing: Techniques and Applications (pp. 605-610). IEEE.

Pearl, J. (2000). Models, reasoning and inference. Cambridge, UK: Cambridge-UniversityPress, 19(2).

Piper, J., & Granum, E. (1989). On fully automatic feature measurement for banded chromosome classification. Cytometry: The Journal of the International Society for Analytical Cytology, 10(3), 242-255.

Ritter, G., & Gallegos, M. T. (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. Pattern recognition letters, 18(6), 525-539.

Steup, M. (2007). The analysis of knowledge. Stanford encyclopedia of philosophy.

Øhrstrøm, P., & Hasle, P. (2007). Temporal logic: From ancient ideas to artificial intelligence (Vol. 57). Springer Science & Business Media.