An Effective Disease Risk Indicator Tool

Kamal Taha
Department of Electrical and Computer Engineering
Khalifa University, Abu Dhabi, UAE
kamal.taha@ku.ac.ae

Paul D. Yoo
Department of Computer Science and Information Systems
Birkbeck College, University of London, UK
paul@dcs.bbk.ac.uk

***Abstract*—Each mixture of deficient molecular families of a specific disease induces the disease at a different time frame in the future. Based on this, we propose a novel methodology for personalizing a person's level of future susceptibility to a specific disease by inferring the mixture of his/her molecular families, whose combined deficiencies is likely to induce the disease. We implemented the methodology in a working system called DRIT, which consists of the following components: logic inferencer, information extractor, risk indicator, and interrelationship between molecular families modeler. The information extractor takes advantage of the exponential increase of biomedical literature to extract the common biomarkers that test positive among most patients with a specific disease. The logic inferencer transforms the hierarchical interrelationships between the molecular families of a disease into rule-based specifications. The interrelationship between molecular families modeler models the hierarchical interrelationships between the molecular families, whose biomarkers were extracted by the information extractor. It employs the specification rules and the inference rules for predicate logic to infer as many as possible probable deficient molecular families for a person based on his/her few molecular families, whose biomarkers tested positive by medical screening. The risk indicator outputs a risk indicator value that reflects a person's level of future susceptibility to the disease. We evaluated DRIT by comparing it experimentally with a comparable method. Results revealed marked improvement.**

Keywords- Gene-disease association, disease risk indicator, information extraction, predicate logic, inference rules.

I. INTRODUCTION

We introduce in this paper a novel methodology for personalizing a person's level of future susceptibility to a specific disease. The methodology overcomes the limitations of current methods. We implemented the methodology in a working system called **D**isease **R**isk **I**ndicator **T**ool (**DRIT**). The proposed system DRIT is able to predict the level of future susceptibility to a specific disease for a person. It is composed of the following four components: information extractor, interrelationship between molecular families modeler, logic inferencer, and risk indicator. The information extractor extracts from biomedical literature the common biomarkers that test positive among most patients with a specific disease. The component employs novel *strict* rule-based information extraction techniques constructed based on established linguistic theories. These strict rules ensure that *only* the biomarkers terms that are closely associated with a disease are extracted.

The interrelationship between molecular families modeler models the hierarchical interrelationships between the molecular families of a disease**,** whose biomarkers were extracted by the information extractor. This helps in inferring the mixture of molecular families, whose combined deficiencies may induce the disease. It also helps in inferring a person's probable deficient molecular families of a disease based on his/her biomarkers that tested positive by medical screening**.**

The logic inferencer infers as many as possible probable deficient molecular families of a disease for a person based on his/her few molecular families, whose biomarkers tested positive by medical screening. This is crucial because, the more deficient molecular families of a disease inferred for a person, the more accurate is the prediction of his/her level of future susceptibility to the disease. With reference to the hierarchical interrelationships between the molecular families, the component first composes rule-based specifications that reflects the relationships between the molecular families of a specific disease. Then, the component uses a person's *initial* deficient molecular families as given premises to recursively trigger the appropriate specification rules by applying the *standard inference rules* of predicate logic. This leads to inferring as many as possible deficient molecular families of a disease for the person. Each *mixture* of molecular families, whose combined deficiencies may induce a specific disease, gives a different indication of future level of susceptibility to the disease [3]. Based on this, the risk indicator component assigns a risk indicator value for a person's level of future susceptibility to the disease based on his/her inferred mixture of deficient molecular families.

II. OUTLINE OF THE APPROACH

Fig. 1 presents the system architecture. It shows the relationships between the four components comprising our proposed system DRIT. With reference to the system architecture in Fig. 1, we outline below the sequential processing steps taken by DRIT to predict the level of future susceptibility to a specific disease for a person:

1. *Information extractor component:* This component extracts from biomedical literature the common Molecular Markers **(MMs)** that test positive among most patients with a specific disease. Section III describes this process in details.
2. *Interrelationship between Molecular Families **(MFs)** modeler component:* This component models the hierarchical interrelationships between the molecular families of a disease, whose MMs were extracted by the information extractor component. The component performs the modeling through the following steps:
   a) *Constructing Molecular Characteristic Trees **(MCTs):*** The component constructs MCTs for each set *S* of biomarkers received from the information extractor that belongs to a same molecular family. Each tree is rooted at one of the biomarkers in the set *S*. Section

IV-A describes this process in details.

b) *Constructing MF* Interrelationships Network ***(MFIN):*** The component constructs a MFIN representing the hierarchical interrelationships between the MFs of the disease based on their shared biological characteristics manifested in their MCTs. Section IV-B describes this process.

3. *Logic inferencer component:* This component applies the inference rules for predicate logic to infer as many as possible deficient MFs for a person. It performs the inferencing through the following two steps:
   a) *Composing rule-based specifications:* The component composes specification rules that reflect the interrelationships between the different MFs of a disease. It composes these rules with reference to the MFIN. Section V-A describes this process.
   b) *Applying the inference rules for predicate logic:* This component uses the person's initial molecular families, whose biomarkers tested positive by medical screening, as given premises to recursively trigger the appropriate specification rules. It does so by applying the *standard inference rules* for predicate logic. Section V-B describes this process.
4. *Risk indicator component:* Based on the mixture of deficient molecular families of the person inferred by the logic inferencer component, this component outputs a risk indicator value. The indicator reflects the person's level of future susceptibility to the disease.

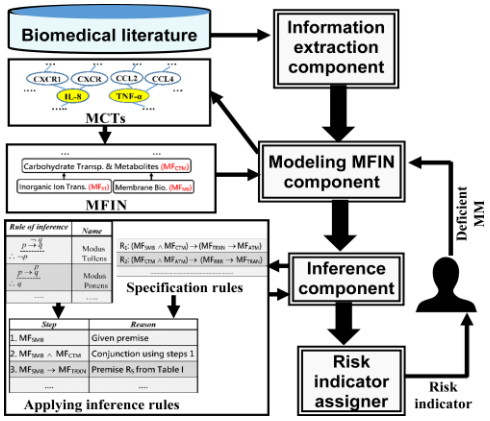

Fig. 1: DRIT system architecture

III. Information Extractor

We first retrieve the biomedical literature associated with a specific disease from reputable biological databases. DRIT extracts from each set of publications associated with a disease the MM terms that are semantically related to the disease terms. We retrieved the literature and disease data from the following databases: (1) PubMed [8] for downloading published works about diseases, (2) Online Mendelian Inheritance in Man (OMIM) [1] for retrieving human genes, genetic disorders, and traits, (3) Human Protein Atlas (HPA) [8] for retrieving expression profiles of human protein coding genes, and (4) UniProtKB [3] for retrieving the functional information on proteins. DRIT employs novel computational linguistic techniques for extracting the MM terms that are semantically related to a disease term. The techniques consider not only the explicit co-occurrences of terms but also their implicit co-occurrences in sentences.

IV. Interrelationship Between MFs Modeler

A. Constructing MCTs

Most molecules associated with a disease have overlapping biological characteristics. To account for these characteristics, we build Molecular Characteristic Trees **(MCTs)** for each MF of a specific disease. An MCT models the hierarchical interrelationships between the molecules of a MF based on their overlapping biological characteristics. A set of MCTs are constructed for each MF. The number of these MCTs is the number of the MMs extracted by the information extractor component (recall Section III) that belongs to the MF. Each MCT will be rooted at a node representing one of the MMs of the MF.

Let S be a set of MFs of a specific disease, whose MMs were extracted by the information extractor component. To account for the common biological characteristics among the molecules of each MF $mf \in S$, we construct MCTs for mf. Each MCT *mct* that belongs to *mf* is constructed as follows. *mct* will be rooted at a node n_i representing a MM $mm \in mf$. Each molecule *mol* that is biologically related to *mm* is represented by a node n_j and is connected to n_i by an edge. The molecules biologically related to *mol* are represented by nodes, which will be connected to n_j by edges. This process continues until all molecules belong to *mf* are exhausted.

Example 1 (running example): Consider the MF "CXC chemokine", which is involved in Type 2 Diabetes **(T2D)**. Fig. 2 shows a fragment of the MCTs for "CXC chemokine". The two MCTs in the figure are rooted at the MMs "IL-8" and "TNF- α", which belong to "CXC chemokine" and involved in T2D. The figure shows fragments of the interrelationships between some of the molecules related to the two MMs.

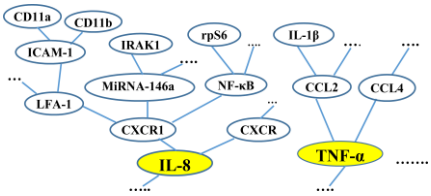

Fig 2: Fragment of MCTs for the MF "CXC chemokine" associated with T2D

B. Constructing MFIN

To infer the MFs, whose combined deficiencies induces a specific disease, we need to identify their interrelationships. These interrelationships will be transformed by DRIT into inference specification rules, which will be used by the system to infer as many as possible deficient MFs for a person. Towards this, we construct a network representing the hierarchical interrelationships between the MFs of a disease based on their shared molecules manifested in the MCTs of these MFs. We call the resulting network MF Interrelationships Network **(MFIN)**.

The hierarchical relationship between two MFs mf_x and mf_y is depicted in the MFIN based on the relative hierarchical levels of their common molecules at their MCTs. Let $\hat{s}$ be a set of common molecules between mf_x and mf_y. Let the hierarchical levels of $\hat{s}$ in the MCT of mf_x be higher than the hierarchical levels of $\hat{s}$ in the MCT of mf_y. In this case, mf_x is more specific and the direction of the relationship between mf_x and mf_y is manifested in the MFIN by an edge from mf_x to mf_y.

Example 2 (running example): Fig. 3 shows a fragment of MFIN depicting the interrelationships between the MFs associated with T2D in our running example.

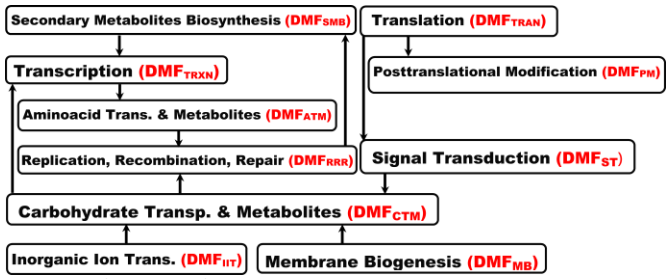

Fig 3: A fragment of MFIN depicting the hierarchical interrelationships between MFs associated with T2D in our running example. MF_{xyz} denotes the MF, whose name abbreviation is xyz

V. Logic Inferencer

A. Composing Rule-Based Specifications

We compose rule-based specifications that reflect the interrelationships between MFs, whose combined deficiencies may induce a specific disease. Eventually, these rules will be used by DRIT as inference rules to infer as many as possible probable deficient MFs of a disease for a person. We composed these rules with reference to the MFIN (recall Section IV-B) that depicts the interrelationships between MFs. Towards this, we convert the interrelationships between the MFs manifested in MFIN into transformation rules. Specifically, we convert the hierarchical interrelationships between the MFs in MFIN by chaining them together into logical transformation rules.

We compose the rule-based specifications in a format resemble the premises of predicate logic [9, 12]. A predicate is a logical statement composed of one or more variables. It is transformed to a proposition by connecting its statements by logical connectives. In the framework of DRIT, the specification rules are developed in the same manner. Specification rules are updated periodically to reflect newly discovered MMs for a disease or/and newly published works about the disease.

Example 3 (running example): Fig. 4 shows a fragment of specification rules that reflect the interrelationships between MFs associated with T2D constructed with reference to the MFIN in our running example shown in Fig. 3.

R_1: $(DMF_{SMB} \wedge DMF_{CTM}) \rightarrow (DMF_{TRXN} \rightarrow DMF_{ATM})$
R_2: $(DMF_{CTM} \wedge DMF_{ATM}) \rightarrow (DMF_{RRR} \rightarrow DMF_{TRAN})$
R_3: $DMF_{RRR} \rightarrow (DMF_{TRAN} \rightarrow (DMF_{PM} \vee DMF_{ST}))$
R_4: $(DMF_{IIT} \vee DMF_{MB}) \rightarrow DMF_{CTM}$
R_5: $DMF_{SMB} \rightarrow DMF_{TRXN}$
R_6: $((DMF_{IIT} \wedge DMF_{MB}) \vee DMF_{TRAN}) \rightarrow DMF_{ST}$
R_7: $DMF_{CTM} \rightarrow DMF_{RRR}$
R_8: $DMF_{TRAN} \rightarrow (DMF_{PM} \vee DMF_{ST})$
.....

Fig. 4: A Sample of specification rules that reflect the interrelationships between MFs associated with T2D constructed with reference to the MFIN in Fig. 3. R_i denotes rule/premise number *i*. The logic symbols "$\wedge$", "$\vee$ ", and "$\rightarrow$" denote conjunction, logical disjunction, and implies respectively

B. Applying the Inference Rules for Predicate Logic

The more deficient MFs of a disease identified for a person, the more accurate is the prediction of his/her level of future susceptibility to the disease. Therefore, we propose to use the inference rules of predict logic [2, 6] to infer as many as possible probable deficient MFs of a disease for a person based on his/her few MFs, whose MMs tested positive by medical screening.

By matching a person's biological molecules (e.g., MMs) that revealed abnormalities for a specific disease by medical screening with the corresponding ones in the MCTs of the disease's MFs, DRIT is able to identify the person's *initial* deficient MFs. DRIT will use these *initial* MFs as given premises to trigger the appropriate specification rules (recall Section V-A) by applying the *standard inference rules* for predicate logic. This will lead to implicitly infer as many as possible probable deficient MFs of the disease for the person. Fig. 5 shows the major standard inference rules for predicate logic [12]. Thus, DRIT employs the following for inferring most of the deficient MFs of a disease for a person: (1) the specification rules (i.e., premises) of a disease, (2) the initial deficient MFs *(i.e., given premises)* for a person identified by medical screening, and (3) the standard inference rules for predicate logic (recall Fig. 5).

The specification rules are triggered by applying the standard inference rules for predicate logic. DRIT triggers recursively the specification rules using the given premises, auxiliary inferred premises, and the standard inference rules for predicate logic. At each recursion, a specification rule (i.e., a premise) is triggered and applied to the premises that have been proven previously. This will lead to a newly proven premise. The conclusions will be a set of inferred MFs. The conclusions are valid, if they have been deducted from all previous premises [11, 12].

Rule of inference	*Name*	*Rule of inference*	*Name*
$\neg q$ $p \rightarrow q$ --- $\therefore \neg p$	Modus Tollens	p --- $\therefore p \vee q$	Disjunctive Amplification
p $p \rightarrow q$ --- $\therefore q$	Modus Ponens	$\neg p \rightarrow$ False --- $\therefore p$	Contradiction
$p \wedge q$ --- $\therefore p$	Simplification	$p \wedge q$ $p \rightarrow (q \rightarrow r)$ --- $\therefore r$	Conditional Proof
p q --- $\therefore p \wedge q$	Conjunction	$p \rightarrow r$ $q \rightarrow r$ --- $\therefore (p \vee q) \rightarrow r$	Proof by Cases
$p \vee q$ $\neg p$ --- $\therefore q$	Disjunctive Syllogism	$p \rightarrow q$ $q \rightarrow r$ --- $\therefore p \rightarrow r$	Law of Syllogism

Fig. 5: Major standard inference rules for predicate logic

Example 4 (running example): Consider that the initial deficient MFs of T2D identified by medical screening for a person are MF_{SMB} and MF_{CTM}. As Fig. 6 shows, the inference rules could infer the following four MFs for the person: (1) MF_{TRXN} (from step 5), (2) MF_{ATM} (from step 8), (3) MF_{RRR} (from step 10), and (4) MF_{TRAN} (from step 13).

Step	*Reason*
1. DMF_{SMB}	Given premise
2. DMF_{CTM}	Given premise
3. $DMF_{SMB} \wedge DMF_{CTM}$	Conjunction using steps 1 and 2
4. $DMF_{SMB} \rightarrow DMF_{TRXN}$	Premise R_5 from Table I
5. DMF_{TRXN}	Modus Ponens using steps 1 and 4
6. $(DMF_{SMB} \wedge DMF_{CTM}) \wedge DMF_{TRXN}$	Conjunction using steps 3 and 5
7. $(DMF_{SMB} \wedge DMF_{CTM}) \rightarrow (DMF_{TRXN} \rightarrow DMF_{ATM})$	Premise R_1 from Table I
8. DMF_{ATM}	Conditional Proof using steps 6 and 7
9. $DMF_{CTM} \rightarrow DMF_{RRR}$	Premise R_7 from Table I
10. DMF_{RRR}	Modus Ponens using steps 2 and 9
11. $DMF_{CTM} \wedge DMF_{ATM} \wedge DMF_{RRR}$	Conjunction using steps 2 and 8 and 10
12. $(DMF_{CTM} \wedge DMF_{ATM}) \rightarrow (DMF_{RRR} \rightarrow DMF_{TRAN})$	Premise R_2 from Table I
13. DMF_{TRAN}	Conditional Proof using steps 11 and 12

Fig 6: Inferring MF_{TRXN}, MF_{ATM}, MF_{RRR}, and MF_{TRAN} from the given premises MF_{SMB} and MF_{CTM}, which are associated with T2D, as described in our running example 4.

VI. Risk Indicator

Each different *mixture* of MFs, whose combined deficiencies may induce a specific disease, gives a different indication for future level of susceptibility to the disease [3]. That is, each different mixture of deficient MFs induces the disease at a different time frame in the future. Thus, a mixture of inferred deficient MFs of a disease for a person can be an indicative of the person's level of future susceptibility to the disease. This led us to assign a risk indicator (e.g., in a scale from 1 to 10) for each mixture of deficient MFs of a specific disease. Each indicator reflects a person's level of future susceptibility to a disease. An indicator value is assigned to a mixture of MFs of a disease based on established and well-known facts about the disease. We collected these facts from the following:

(1) Information extracted from biomedical literature associated with each disease.

(2) Information obtained by consulting medical professionals. We compiled a table of risk indicator values and their corresponding mixtures of MFs of a disease. By matching a person's deficient mixture of MFs with the different mixtures in the table, DRIT will return the corresponding risk indicator in the table.

VII. Experimental Results

We implemented DRIT in Java and ran it under Windows 10 Pro and Intel(R) Core(TM) i7-6820HQ processor. The RAM and CPU of the machine have 16 GB and 2.70 GHz respectively. We evaluated DRIT by comparing it experimentally with SCAIView [13]. SCAIView incorporates the following two external software components for retrieving biomedical literatures associated with biomarkers: (1) ProMiner, and (2) SCAIView [1]. Retrieved biomedical texts are ranked based on the frequency of cooccurrences of biomarker-disease associations included within them. Extracted biomarker-disease associations are organized into classes. We used UMLS [14] database for constructing a disease dictionary. We selected the concepts associated with the semantic type "Disease". We compiled the terminology of genes' names by cross-referencing and integrating data extracted from the following: (1) UniProt [3], (2) HGNC [5], and (3) NCBI-Gene [6] databases. We then used the MeSH Browser [1] to map the extracted genes and diseases terms to MeSH IDs. Finally, we retrieved the biomedical literature associated with the extracted genes and diseases terms from PubMed by submitting the following PubMed query: ("diseases" [MeSH Terms]) AND "genes" [MeSH Terms] AND (has abstract [text]) AND (English [lang]) AND ("0001/01/01" [PDAT]: "2019/04/31" [PDAT]). This resulted in 403,742 publications. We evaluated DRIT by comparing it with SCAIView [13]. We ran DRIT and SCAIView against the retrieved biomedical literature described previously. DRIT identified 3,418 gene-disease associations. We compared the gene-disease associations identified by DRIT and SCAIView with the corresponding gold standard ones that we retrieved from BIOBASE database [4]. We evaluated the prediction accuracy of the two systems in terms of Recall, precision, and F-value, where Recall = TP/(TP+FN), Precision = TP/(TP+FP), F-value = (2 Precision * Recall)/(Precision + Recall), FP = False positive, TP = True Positive, and FN = False Negative. Fig. 7 plots the results. The results revealed that DRIT outperformed SCAIView. It revealed that the performance of DRIT over SCAIView kept increasing at a higher rate as the size of dataset kept being increased. This is advantageous to DRIT, since the size of biomedical literature associated with MMs increases constantly over time in real-world setting. The results revealed also that the strict linguistic rules employed by DRIT contributed to its performance.

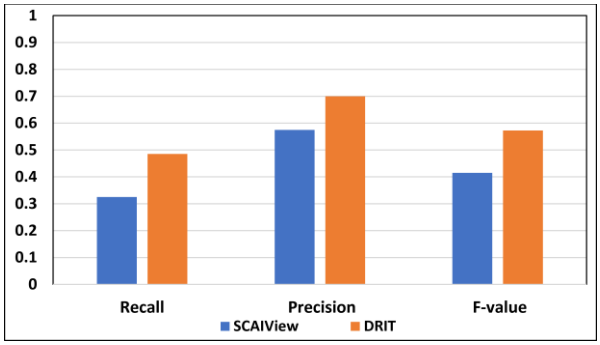

Fig. 7: The overall average Recall, Precision, and F-value of DRIT and SCAIView for predicting gene-disease associations.

References

[1] Amberger,.S. et al. (2015). "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders". Nucleic Acids Res. 43: D789–98.

[2] Al-Aamri, A., Taha, K., Homouz, D., Al-Hammadi, Y., and Maalouf, M. "Analyzing a co-occurrence gene-interaction network to identify disease-gene association", *BMC Bioinformatics 20*, Article number: 70, 2019.

[3] Bairoch A, et al: "The Universal Protein Resource (UniProt)". Nucleic Acids Research 2005, 33(1):154-159.

[4] BIOBASE database: https://www.qiagenbioinformatics.com/

[5] HGNC, http://www.genenames.org/cgi-bin/hgnc_downloads.

[6] Kenneth H. Rosen: Discrete Mathematics and its Applications, Fifth Edition, McGraw-Hill, 2003.

[7] NCBI-Gene, ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz.

[8] PubMed. Available at: https://www.ncbi.nlm.nih.gov/pubmed/

[9] Taha, K. "Employing the Inference Rules of Predicate Logic for Predicting Protein Functions". 41th IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Chicago, USA, May 2019

[10] The Human Protein Atlas. Available at: www.proteinatlas.org.

[11] Taha, K., Iraqi, Y., and Al-Aamri, A. "Predicting Protein Functions by Applying Predicate Logic to Biomedical Literature", BMC Bioinformatics 20, Article number: 71, February 2019.

[12] Warner, R.M. (2013) Applied Statistics: From Bivariate through Multivariate Techniques. SAGE Publications, Thousand Oaks.

[13] Younesi, E., et al., Mining biomarker information in biomedical literature. BMC medical informatics and decision making, 2012. 12(1): p. 148.

[14] 2019AA Full UMLS Release Files, https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.