Conditional Random Fields-based Approach to Classification: Application to Life Sciences

Grbić, Milana

Abstract: ***Data prediction is one of the most challenging problems in information sciences and has a lot of applications in various fields of science. In the field of biochemistry, predicting the role of metabolite can be of great importance for obtaining additional information about processes within the organisms, as well as for a better understanding of the metabolic pathways. Due to numerous studies based on energy consumption or production, it is useful to determine whether metabolites participate in production or usage of energy in these reactions. i.e. to classify them as energy producers or consumers.***

In this paper Conditional Random Fields (CRF) method is used for structural prediction of roles the metabolites play in metabolic reactions. For prediction of metabolite roles, three different sets of feature functions, which involve information about elements in the nearest neighborhood and information about their labels, are proposed. By using the CRF++ software package, the proposed CRF method is tested on a set of metabolite reactions from the yeast Saccharomyces cerevisiae. In addition, the results of the classification are verified by another CRF implementation (CRFsuite) from the literature. Obtained results indicate a high level of accuracy of the proposed CRF approach.

Index Terms: ***conditional random fields, data classification, prediction, metabolite reactions***

1. Introduction

Task of data prediction is one of the most important problems in information sciences. A lot of applications can be found in various fields of science, like prediction whether a received email is spam or not, prediction if a cancer cell is benign or malignant, prediction of protein disorder. Prediction encompasses various types of classification problems - as specific type of prediction problems - such as text classification, classification of the validity of credit cards etc.

The goal of this research is to predict the role of metabolites in metabolic reaction is considered. Based on this prediction, each metabolite is classified in a particular class, which corresponds to its role in a reaction. Prediction of roles of metabolites is important for further understanding of metabolic pathways. Each metabolic pathway can be considered as a series of chemical reactions which involve reactants, products and various intermediates. There are two basic types of biochemical reactions that are characterized by their presence in either anabolic or catabolic pathways, i.e. pathways with the ability to synthesize molecules with the utilization of energy or to degrade metabolites with releasing energy. Adenosine triphosphate (ATP) appears in both types of pathways either as a reactant (anabolic pathway) or as a product (catabolic pathway). In anabolic pathways, ATP can be degraded to adenosine diphosphate (ADP) or to adenosine monophosphate (AMP), releasing one or two phosphate groups, which further may stay free or can be bound with a metabolite in a phosphorylation process. In catabolic pathways, reverse processes of synthesis ATP from its lower forms ADP and AMP occur. Identifying the role of metabolites in reaction could be useful for finding out more about connections between metabolites based on their involvement in the same reaction. Also, obtained results can be used for a deeper analysis of metabolism of the particular organism.

Conditional Random Fields (CRF) is an approach to prediction, which includes the dependency between variables in the prediction process. In this research a linear CRF is adapted and applied to a list of chemical reactions. Adaptation of existing methods is a common methodological approach to innovation in computer science [1]. The list of chemical reactions is considered as a sequence of sentences. The chosen set of reactions belongs to the particular pathways with the ability to synthesize molecules with the utilization of energy or to degrade metabolites with releasing energy. In the CRF method, the context is taken into account. Since role of a metabolite depends on the neighboring metabolites in the reaction as well as on their labels, this method seems to be convenient for such prediction. We believe that the idea of combination of information about neighboring elements and labels which are input for CRF method, can give the more precise prediction of the role of metabolites. To our knowledge, the problem of predicting the metabolite roles has not been considered in such a way. Still, in the literature one can find several existing applications of CRF method and some similar predictions of biological elements.

The paper is organized as follows: the next section reviews work related to applications of the CRF method and work related to prediction of roles biological elements play in different processes. In Section 3 a description of the proposed CRF model for classification of metabolite roles is provided. Experimental results are shown and discussed in Section 4. The last section is the conclusion.

2. *Related Work*

CRF method is widely used in the field of name entity recognition. The task is identifying the classes of interest to which particular word or phrase belongs. Several recent results are reported in papers [2-5]. WebListing technique based on CRF method, presented in [2], builds seeds for lexicons based on labeled data. These seeds are further significantly augmented by using HTML data on the Web. Different approaches of using CRF based methods for recognizing specific biological terms, such as PROTEIN, DNA, RNA, CELL-LINE and CELL-TYPE in biomedical abstracts are described in [3]. A CRF based open source machine-learning system called BANNER is designed to maximize domain independence, also achieving significantly better performances than other baseline systems [4]. Hybrid model LSTM-CRF, which is combination of Long Short-term Memory Networks (LSTM) and CRF, is also used for named entity recognition [5].

A lot of applications in text preprocessing and analysis make use of CRF method. For example, it is used for shallow parsing [6], for analysis of sentence which first recognizes nouns, verbs, adjectives, etc. and after that groups them to higher order terms such as phrases. Problems such as Part Of Speech (POS) and Chunking can also be solved by methods which are based on CRF [7]. Tree Conditional Random Fields is used for semantic role labeling [8]. In [9], CRF for extraction information from tables is presented. CRF method also can be used for word alignment [10], document summarization [11] and interactive question answering [12].

CRF is also a base for some methods for image segmentation. Discriminative Random Fields (DRF) model, which is based on concept of CRF, has an application in modeling spatial dependency [13]. Shape and texture can be jointly modeled into textons, which are novel features incorporated in CRF method used for image segmentation [14]. Foreground and shadow segmentation can be done using a special class of CRF, named dynamic conditional random fields (DCRF), introduced in [15].

In [16] CRF is used as a replacement of heuristic approaches for stereo vision algorithms. A large number of stereo datasets with ground-truth disparities was constructed, and a subset of these datasets was used to learn the parameters of CRF.

CRF has been intensively used in various fields of bioinformatics. In [17], it is used for tagging gene and protein mentions in text. As it is already mentioned, CRF is a suitable method for name entity recognition in biomedical texts [3]. The first comparative gene predictor based on semi-Markov conditional random fields (SMCRFs) named Conard is presented in [18]. One of the most important problems in bioinformatics is prediction of protein folding. In [19], an effective solution based on segmentation conditional random fields (SCRFs) is proposed. Estimation of parameters used for RNA structural alignment and structural alignment search based on CRF are shown better and more accurate than existing methods [20]. Markov Random Fields, where variables are jointly Gaussian, is called Gaussian CRF (GCRFs). It has an application in computer vision [21]. Extension of the Gaussian CRF, called Directed Gaussian conditional random fields (DirGCRF), is introduced to allow modeling asymmetric relationships [22].

Predicting roles of biological elements in different processes was analyzed in several papers. In [23], the authors considered the role of metabolites in predicting drug-drug interactions. Focus of that research was on irreversible inhibition of cytochrome P450 enzymes and on the metabolites which are fundamental for that inhibition. By using the Bayesian approach and Labelled Multilevel Neighborhoods of Atoms (LMNA) descriptors, the prediction of reacting atom(s) in molecules was carried out in [24]. The method for prediction to which metabolic pathway a particular compound belongs was described in [25]. That method is based on the highest interaction confidence scores. The Random Forest based approach for prediction of metabolic enzymes and gut bacteria was presented in [26].

3. *CRF Method for Predicting Metabolite Roles*

3.1. Problem Definition

Let a list of metabolic reactions be given. Each reaction contains several metabolites, which appear at the left-hand and the right-hand sides of the arrow (see two examples of metabolite reactions in Figure 1). Metabolites from the left-hand side are usually called reactants and the metabolites from the right-hand side of the arrow are products of the reaction. The classification problem can be formulated as follows: given a reaction, assign a label to every metabolite, representing the role the metabolite plays in the reaction. The set of labels are given in advance and in general, it can be based on a particular need in a considered context.

One possible labeling can be based on metabolite's involvement in the energy transfer or in the phosphorylation process. In the approach presented in this paper, the following eight classes are identified:

- Label 1: "*donor of one or two phosphate group(s)"*
- Label 2: *"acceptor of one or two phosphate group(s)"*
- Label 3: *"free phosphate group(s)"*
- Label 4: *"phosphate (compound built by binding one or two phosphate groups(s) with metabolite)"*
- Label 5: *"lower forms of ATP in anabolic pathways"*
- Label 6: *"ATP synthetized in catabolic pathways"*
- Label 7: *"phosphate group which binds in catabolic pathways"*
- Label 8: *"the class containing the rest of metabolites"*

A reaction is considered as a sequence of several elements which are either on its left-hand or on its right-hand side. Therefore, it is important to recognize the arrow as a mark which separates these two sides of the reaction. So, an additional label, Label 9: *"arrow"*, is introduced, only for this purpose.

It should be noted that if some other lists of reactions are considered, the list of labels can be extended, reduced or modified.

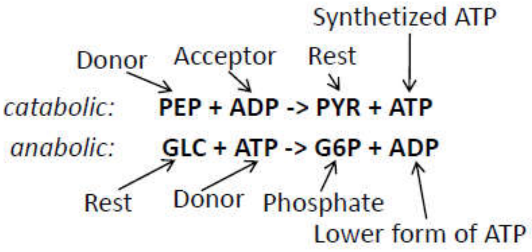


Figure 1: Example of labeling metabolite in reactions

Example 1:

In Figure 1 two reactions are shown. Let us consider the first reaction. This reaction is a part of catabolic pathway. PEP or phosphoenol-pyruvate is an important compound, which contains high energy phosphate bonding [27]. In this reaction PEP is "a donor of one phosphate group". The other reactant is adenosine diphosphate (ADP) which consists of three components: adenine, sugar backbone and two phosphate groups [28]. In this reaction, it is "an acceptor of phosphate group" which is released by PEP. ADP binds that phosphate group and forms adenosine triphosphate (ATP), so the product ATP is labeled as Label 6, i.e. "ATP synthetized in this catabolic pathway". The second reaction from Figure 1 is a part of anabolic pathway. So, ATP is "a donor of one phosphate group" and ADP is "a lower form of ATP". G6P is "a phosphate (compound built by binding one or two phosphate groups(s) with metabolite)" and GLC belongs to "the class of the rest of metabolites".

3.2. *CRF Method for Metabolite Classification*

CRF is a discriminative probabilistic model of machine learning for a structured prediction. Structured prediction refers to supervised machine learning technique that involves predicting structured objects, such as sequence, graph, tree. CRF is introduced in [7] and its main principle can be explained on the problem of labeling a sequence of data. Let T be the length of the sequence. Let $\mathbf{x} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_T}\}$ represent characteristics of the elements of the sequence, where every $\mathbf{x}_i$ is a vector of characteristics of the element at the position i - x_i. For every element x_i the task is to find the appropriate label y_i based on vector $\mathbf{x}_i$ and neighboring labels. For solving this problem the training set of data $(\mathbf{x}_i, y_i)$ with correct information about labels is provided. The problem of prediction is actually the problem of finding the probability $p(y|\mathbf{x})$, where $y = \{y_1, y_2, \dots, y_T\}$. This probability can be determined in two ways:

i. from the joint probability $p(y, \mathbf{x})$ and the probability $p(\mathbf{x})$ – *generative approach;*
ii. directly, by modeling this conditional probability – *discriminative approach.*

For every generative method there exist a discriminative analogue and vice versa [29]. As it has been already mentioned, the CRF belongs to the class of discriminative methods, i.e. the probability $p(y|\mathbf{x})$ is directly calculated. It is well known that Hidden Markov Models (HMM) is the generative analogue of CRF.

CRF usually imposes the following calculations:

$$p(y|x) = \frac{1}{Z(x)} exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k\, f_k(y_{t-1}\,, y_t\,, \mathbf{x_t})\right\},$$

where f_k represents feature function and θ_k are the components of parameter vector, for $1 \le k \le$

K. The normalization factor is defined by

$$Z(x) = \sum_{y \in Y^T} exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, \mathbf{x_t})\right\}.$$

Let us notice that the feature functions $f_k(y_{t-1}, y_t, \mathbf{x_t})$ use the vector $\mathbf{x_t}$ as an argument, which indicates that all the components of global observations $\mathbf{x}$, that are needed for computing features at the position t, are available. For example, if the next element x_{t+1} is used as a feature for CRF, it is assumed that the information about the identity of that element is included in the vector $\mathbf{x_t}$ [30].

Chemical reactions can be considered as sentences. Although, there are no strict rules for writing reactions, some common conventions still exist. For example, a donor is usually written at the first position or before an acceptor, lower forms of ATP are usually written at the last position or at the position next to the last in the reaction. Therefore, for determining the role of a particular metabolite, it is justified to consider its neighbors and their labels. To examine this hypothesis, three different feature models (named A, B and C), using different sets of feature functions, are constructed.

Model A

In the first model, feature functions are based on the information about the nearest metabolites in the reaction and the information about the label of current element. More precisely, for determining the label of a metabolite at the position t in a reaction, the information about the nearest metabolite as well as the information about nearest consecutive (left and right) pairwise metabolites are taken into account. To formalize, the information about the metabolites at positions from the set $\{t-2,\ t-1,\ t,\ t+1, t+2\}$ and information about the label of the current element at the position t are taken into consideration. The following are the feature functions used:

$$f_{x,y}^{-2}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t-2} = x, y_t = y),$$

$$f_{x,y}^{-1}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t-1} = x, y_t = y),$$

$$f_{x,y}^{0}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_t = x, y_t = y),$$

$$f_{x,y}^{+1}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t+1} = x, y_t = y),$$

$$f_{x,y}^{+2}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t+2} = x, y_t = y).$$

$$f_{x,x',y}^{-1,0}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t-1} = x,\ x_t = x', y_t = y),$$

$$f_{x,x',y}^{0,+1}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_t = x,\ x_{t+1} = x', y_t = y).$$

Let us explain the notation of these functions. In all formulas, I is the indicator function, i.e. it is equal to 1 if the condition is valid, otherwise 0. In the superscript of a function f, the mark $-i$, (respectively $+i$) $i \in \{1,2\}$, means that the information about the metabolite, which is on the distance i to the left (respectively to the right) from the considered one, is taken into account. The mark 0 means that the information about the current element is considered.

Thus the set of feature functions used in Model A is

$$F_A = \left\{f_{x,y}^{-2}, f_{x,y}^{-1}, f_{x,y}^{0,}, f_{x,y}^{+1}, f_{x,y}^{+2}, f_{x,x',y}^{-1,0}, f_{x,x',y}^{0,+1} | x, x' \in X, y \in Y\right\},$$

where X is the set of all elements which are present in given reactions and Y is the set of all possible labels.

Model B

The first five functions from the model A are also used in the model B. This subset of feature functions is extended by the following functions, which use information about the label of the previous metabolite in the reaction and information about the nearest metabolites:

$$g_{x',x,y',y}^{-1,0}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t-1} = x', x_t = x,\ y_{t-1} = y', y_t = y\ ),$$

$$g_{x',x,y',y}^{0,+1}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_t = x', x_{t+1} = x, y_{t-1} = y',\ y_t = y),$$

$$g_{x'',x',x,y',y}^{-2,-1,0}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t-2} = x'', x_{t-1} = x', x_t = x, y_{t-1} = y', y_t = y),$$

$$g_{x'',x',x,y',y}^{-1,0,+1}(y_{t-1}, y_t, \mathbf{x_t}) = I(x_{t-1} = x'', x_t = x', x_{t+1} = x, y_{t-1} = y', y_t = y),$$

$$g_{x'',x',x,y',y}^{0,+1,+2}(y_{t-1}, y_{t}, \mathbf{x_t}) = I(x_t = x'', x_{t+1} = x', x_{t+2} = x, y_{t-1} = y', y_t = y).$$

So, the set of feature functions in Model B is

$$F_B = \{f_{x,y}^{-2}, f_{x,y}^{-1}, f_{x,y}^{0,}, f_{x,y}^{+1}, f_{x,y}^{+2}, g_{x',x,y',y}^{-1,0}, g_{x',x,y',y}^{0,+1}, g_{x'',x',x,y',y}^{-2,-1,0}, g_{x'',x',x,y',y}^{-1,0,+1}, g_{x'',x',x,y',y}^{0,+1,+2} | x, x' \in X, y, y' \in Y\}$$

The function $g_{x'',x',x,y}^{-2,-1,0}$ takes into account information about the metabolites at positions $t-2,\ t-1$ and t. Similarly, the function $g_{x'',x',x,y}^{0,+1,+2}$ records information about metabolites at positions $t,\ t+1$ and $t+2$, while the function $g_{x'',x',x,y}^{-1,0,+1}$ records information about metabolites at positions $t-1,\ t$ and $t+1$. All functions take into account the information about the labels of the previous element and the current one.

Model C

The set of feature function in the model C contains only functions which take into account information about the label of the previous element. This set can be derived from the set F_B

by omitting the functions from the model A, so

$$F_C = F_B \setminus \{f_{x,y}^{-2}, f_{x,y}^{-1}, f_{x,y}^{0}, f_{x,y}^{+1}, f_{x,y}^{+2} | x \in X, y \in Y\}.$$

Example 2:

In order to illustrate how the proposed feature functions are calculated, let us consider the reaction:

ATP + AC + COA → AMP + PPI + ACCOA.

The reactants and products are labeled as follows:

1. ATP – Label 1: "*donor of one or two phosphate group(s)"*
2. AC – Label 8: *"the class containing the rest of metabolites"*
3. COA – Label 8: *"the class containing the rest of metabolites"*
4. → - Label 9: *"arrow"*
5. AMP – Label 5: *"lower forms of ATP in anabolic pathways"*
6. PPI – Label 3: *"free phosphate group(s)"*
7. ACCOA – Label 8: *"the class containing the rest of metabolites"*

Let each metabolite be assigned its position in the reaction. Let us, for example, take $t = 3$, (x_3 = COA) and then the values of considered functions are

$$f_{x,y}^{-2} = \begin{cases} 1, x = \text{ATP and } \; y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$f_{x,y}^{-1} = \begin{cases} 1, x = AC \text{ and } y = Label \text{ 8} \\ \quad 0 \text{ , otherwise} \end{cases}$$

$$f_{x,y}^{0} = \begin{cases} 1, x = \text{COA and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$f_{x,y}^{+1} = \begin{cases} 1, x = \rightarrow \text{ and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$f_{x,y}^{+2} = \begin{cases} 1, x = \text{AMP and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$f_{x,x',y}^{-1,0} = \begin{cases} 1, x = \text{AC}, x' = \text{COA and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$f_{x,x',y}^{0,1} = \begin{cases} 1, x = \text{COA}, x' = \rightarrow \text{ and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$g_{x',x,y',y}^{-1,0} = \begin{cases} 1, x' = \text{AC}, x = \text{COA}, \text{y}' = \text{Label 8 and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

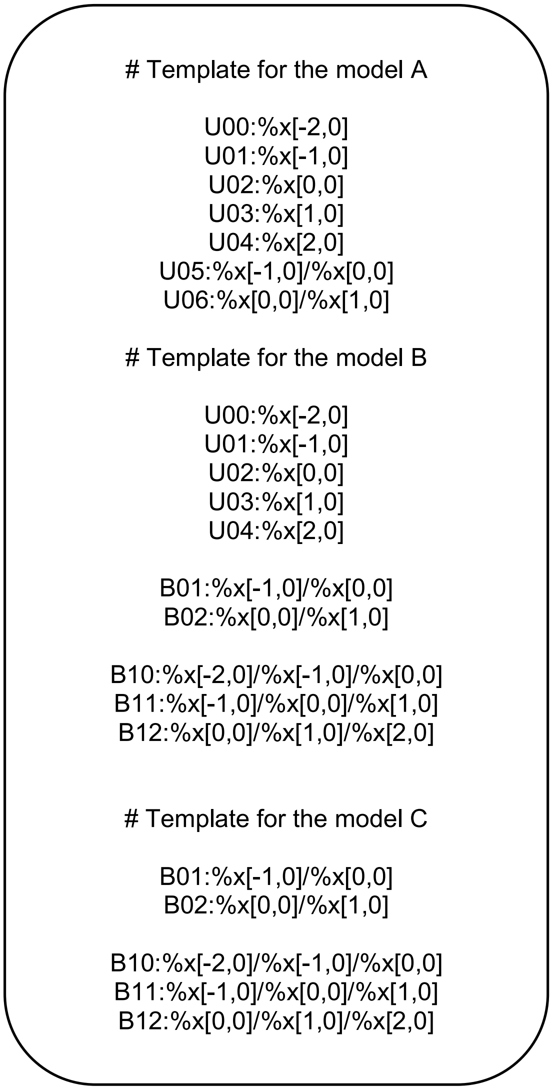
$$g_{x',x,y',y}^{0,+1} = \begin{cases} 1, x' = \text{COA}, x = \rightarrow, \; \text{y}' = \text{Label 8 and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$g_{x'',x',x,,y',y}^{-2,-1,0} = \begin{cases} 1, x'' = \text{ATP}, x' = \text{AC}, x = \text{COA}, \text{y}' = \text{Label 8 } \text{ and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$g_{x'',x',x,y',y}^{-1,0,1} = \begin{cases} 1, x'' = \text{AC}, x' = \text{COA}, x = \rightarrow, \text{y}^{'} = \text{Label 8 } \text{ and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}$$

$$g_{x'',x',x,y',y}^{0,1,2} = \begin{cases} 1, x'' = \text{COA}, x' = \rightarrow, x = \text{AMP}, \text{y}' = \text{Label 8 } \text{ and } y = \text{Label 8} \\ \quad 0, \text{otherwise} \end{cases}.$$

The feature functions for other metabolites can be analyzed in a similar way.



```
# Template for the model A

U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

# Template for the model B

U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]

B01:%x[-1,0]/%x[0,0]
B02:%x[0,0]/%x[1,0]

B10:%x[-2,0]/%x[-1,0]/%x[0,0]
B11:%x[-1,0]/%x[0,0]/%x[1,0]
B12:%x[0,0]/%x[1,0]/%x[2,0]


# Template for the model C

B01:%x[-1,0]/%x[0,0]
B02:%x[0,0]/%x[1,0]

B10:%x[-2,0]/%x[-1,0]/%x[0,0]
B11:%x[-1,0]/%x[0,0]/%x[1,0]
B12:%x[0,0]/%x[1,0]/%x[2,0]
```

Figure 2 : An excerpt from the template file

4. Experimental Results and Discussion

This section presents experimental results obtained by the proposed CRF models. All tests are carried out on the Intel i5 @2.5 GHz with 8 GB RAM. In order to implement all proposed models (A, B and C), the software package CRF++ [31] is used. CRF++ allows user-defined templates, so it is suitable for an approach which includes custom feature functions. In this package, calculation of the CRF model is implemented as a forward-backward algorithm and a logarithmic computation is used [32].

In addition, CRF++ allows customization of two other parameters:

- parameter *c*, which is used for obtaining balance between overfitting and underfitting. The default value is set to 1. In our tests, two additional values, 1.5 and 2, were also used.

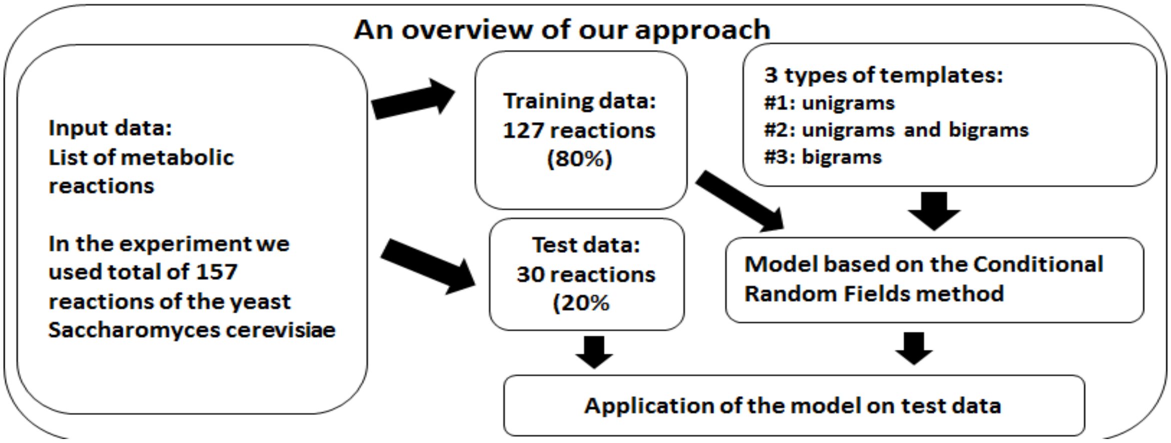

An overview of our approach
Input data:
List of metabolic reactions
In the experiment we used total of 157 reactions of the yeast Saccharomyces cerevisiae
Training data:
127 reactions
(80%)
Test data:
30 reactions
(20%
3 types of templates:
#1: unigrams
#2: unigrams and bigrams
#3: bigrams
Model based on the Conditional Random Fields method
Application of the model on test data

Figure 3: Overall approach

- parameter *f*, which is an integer number that sets the cut-off threshold for the features. Only features whose number of occurrences is equal or greater than *f* are taken into account. The default value is set to 1. Two more values, 2 and 3, were used in the tests.

Dataset used for the tests contains biological information about the metabolism of the yeast Saccharomyces cerevisiae. This dataset consists of a list of 157 metabolic reactions, taken from [33]. These reactions are chosen from the list of all metabolic reactions of this organism - as the reactions which are the part of the energy transfer or the part of the phosphorylation process.

The tests were performed in four phases. In the first phase, the set of the chosen reactions was divided into training data and test data, in ratio 80:20. Templates were formed in the second phase, taking into account the information about the current metabolite in the reaction and information about the other metabolites from the same reaction. As it is described in Subsection 3.2, three different templates were used, corresponding to the models A, B and C. The templates whose name starts with ‘U’ are called unigrams and they take into account information about label of the current element. The templates with the mark ‘B’ (bigrams) use the information about labels of the previous and of the current element. An excerpt of the template file is shown in Figure 2. After the templates were formed, the algorithm entered into the third phase, i.e. the construction of the CRF model, based on the training data and the template file. And finally, in the fourth phase, the model was applied on the test data. Graphical representation of the overall approach is shown in Figure 3.

For a deeper analysis of the proposed CRF models, in a series of tests, the following combinations of control parameters were used:

- default combination, in which *f=1* and *c=1*,
- *f=3* and *c=1.5*,
- *f=2* and *c=1.5*,
- *f=3* and *c=2*,
- *f=2* and *c=2*.

Obtained results are shown in Table 1. In the first column, the name of the model is given. In the rest of the table, for each combination of control parameters and for each model, the accuracy obtained on the test data is presented. The accuracy is calculated in the standard way as ratio between number of correctly classified elements and total number of elements.

From Table 1, it can be seen that models A and B are more accurate than the model C and results of the model A are slightly better than results of the model B. For the model A, the difference between accuracy of obtained results for different parameter combinations is very small, which indicates that the model A is the most stable one. If we consider all three models together, the default parameter combination (last column) gives the best results on average.

Table 1: Experimental results obtained by CRF++

Parameters	*f=3, c=1.5*	*f=2, c=1.5*	*f=3, c=2*	*f=2, c=2*	*Default* *f=1, c=1*
Model					
Model A	93.60465%	94.18605 %	94.18605%	94.18605%	93.02326%
Model B	93.60465%	93.60465%	93.60465%	93.60465%	91.86047%
Model C	66.86047%	68.02326%	66.86047%	68.02326%	77.32558%

Table 2: Performances of the model A with control parameters *f=3* and *c=2*

Class	#match	#model	#ref	precision	recall	F1
Label 1	30	31	30	0.967742	1	0.9836066
Label 2	2	2	2	1	1	1
Label 3	11	11	12	1	0.917	0.9565217
Label 4	11	12	17	0.916667	0.647	0.7586207
Label 5	27	28	27	0.964286	1	0.9818182
Label 6	2	2	2	1	1	1
Label 7	1	1	1	1	1	1
Label 8	47	54	50	0.87037	0.94	0.9038462
Label 9	31	31	31	1	1	1
average				0.968785	0.945	0.9538237

To validate the quality of the proposed models, the same dataset was adapted and tested by another CRF software - CRFsuite. This software reads training data and automatically generates all the necessary state and transition features based on the data [34]. The accuracy obtained by CRFsuite is 94.76% which is very similar to the best results obtained by the proposed model A.

In order to further investigate performance of the proposed models, a comparative analysis of the results obtained for each of the nine label classes, by both CRF packages, was performed. In the case of CRF++, the most successful model and parameter combination, i.e. the model A with control parameters *f=3, c=2* is chosen. For each of 9 classes, three different measures were calculated:

- precision, calculated as

$$\text{precision} = \frac{\#\text{match}}{\#\text{model}}$$

- recall, calculated as

$$\text{recall} = \frac{\#\text{match}}{\#\text{ref}}$$

- F1 measure, calculated as

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

where #match, #model and #ref are the total number of matched labels, the total number of labels in the model and the total number of labels in the test set, respectively.

Results for each label, obtained by the chosen model A tested by CRF++ and the CRFsuite model, are shown in details in Tables 2 and 3, respectively. In both tables data are organized as follows. The first column contains the label of the class. In the next three columns #match, #model and #ref are shown. Last three columns contain information about three measures: precision, recall and F1 measure. In the last row average values for these three measures are shown.

From Tables 2 and 3 it can be seen that both models achieve accurate results for all labels. For each measure, on average, the model A is slightly better than the CRFsuite model. Still, the results shown in these tables indicate that both models perform similarly on the considered data set.

5. Conclusion

Classification of roles of metabolites can be of great importance for further understanding of metabolic process in different organisms. The proposed approach to classification is based on the conditional random field method. Metabolic reactions were considered as sequences of elements, enabling the construction of different feature functions, based on unigrams and/or bigrams. In the paper, three different models were proposed for constructing templates, which are further used for CRF model construction.

Performances of the proposed models were tested on a real set of biological data. The obtained results indicate high accuracy of the models. The detailed analysis in previous section show that the proposed method found appropriate labels for the most of the metabolites, so it indicates that it can be used for solving the considered problem.

Future work may include further application of the proposed models, especially to larger data sets. Another promising extension of this research may involve development of other feature functions containing additional information about characteristics of metabolites.

Table 3: Performances of the CRFsuite software model

	#match	#model	#ref	precision	recall	F1
Label 1	29	30	30	0.966667	0.966667	0.966667
Label 2	2	2	2	1	1	1
Label 3	10	11	12	0.909091	0.833333	0.869565
Label 4	16	20	17	0.8	0.941176	0.864865
Label 5	27	29	27	0.931034	1	0.964286
Label 6	1	1	2	1	0.5	0.666667
Label 7	1	1	1	1	1	1
Label 8	46	47	50	0.978723	0.92	0.948454
Label 9	31	31	31	1	1	1
			average	0.953946	0.906797	0.920056

References

[1] Blagojević, V., Bojić, D., Bojović, M., Cvetanović, M., Đorđević, J., Đurđević, Đ., ... & Milutinović, V. (2017). A systematic approach to generation of new ideas for PhD research in computing. In Advances in computers (Vol. 104, pp. 1-31). Elsevier.

[2] McCallum, A., & Li, W. (2003, May). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 188-191). Association for Computational Linguistics.

[3] Settles, B. (2004, August). Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (pp. 104-107). Association for Computational Linguistics

[4] Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Biocomputing 2008 (pp. 652-663).

[5] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[6] Sha, F., & Pereira, F. (2003, May). Shallow parsing with conditional random fields. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 134-141). Association for Computational Linguistics.

[7] Lafferty, J., McCallum, A. and Pereira, F. CN "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[8] Cohn, T., & Blunsom, P. (2005, June). Semantic role labelling with tree conditional random fields. In Proceedings of the Ninth Conference on Computational Natural Language Learning (pp. 169-172). Association for Computational Linguistics

[9] Pinto, D., McCallum, A., Wei, X., & Croft, W. B. (2003, July). Table extraction using conditional random fields. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 235-242). ACM.

[10] Blunsom, P., & Cohn, T. (2006, July). Discriminative word alignment with conditional random fields. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 65-72). Association for Computational Linguistics.

[11] Shen, D., Sun, J. T., Li, H., Yang, Q., & Chen, Z. (2007, January). Document Summarization Using Conditional Random Fields. In IJCAI (Vol. 7, pp. 2862-2867).

[12] Hickl, A., & Harabagiu, S. (2006, June). Enhanced interactive question-answering with conditional random fields. In Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006 (pp. 25-32). Association for Computational Linguistics.

[13] Kumar, S., & Hebert, M. (2006). Discriminative random fields. International Journal of Computer Vision, 68(2), 179-201.

[14] Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2006, May). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In European conference on computer vision (pp. 1-15). Springer, Berlin, Heidelberg.

[15] Wang, Y., Loe, K. F., & Wu, J. K. (2006). A dynamic conditional random field model for foreground and shadow segmentation. IEEE transactions on pattern analysis and machine intelligence, 28(2), 279-289.

[16] Scharstein, D., & Pal, C. (2007, June). Learning conditional random fields for stereo. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on (pp. 1-8). IEEE.

[17] McDonald, R., & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. BMC bioinformatics, 6(1), S6.

[18] DeCaprio, D., Vinson, J. P., Pearson, M. D., Montgomery, P., Doherty, M., & Galagan, J. E. (2007). Conrad: gene prediction using conditional random fields. Genome research, 17(9), 000-000.

[19] Liu, Y., Carbonell, J., Weigele, P., & Gopalakrishnan, V. (2006). Protein fold recognition using segmentation

conditional random fields (SCRFs). Journal of Computational Biology, 13(2), 394-406.

[20] Sato, K., & Sakakibara, Y. (2005). RNA secondary structural alignment with conditional random fields. Bioinformatics, 21(suppl_2), ii237-ii242.

[21] Tappen, M. F., Liu, C., Adelson, E. H., & Freeman, W. T. (2007, June). Learning gaussian conditional random fields for low-level vision. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[22] Vujicic, T., Glass, J., Zhou, F., & Obradovic, Z. (2017). Gaussian conditional random fields extended for directed graphs. Machine Learning, 106(9-10), 1271-1288.

[23] VandenBrink, B. M., & Isoherranen, N. (2010). The role of metabolites in predicting drug-drug interactions: Focus on irreversible P450 inhibition. Current opinion in drug discovery & development, 13(1), 66.

[24] Rudik, A. V., Dmitriev, A. V., Lagunin, A. A., Filimonov, D. A., & Poroikov, V. V. (2016). Prediction of reacting atoms for the major biotransformation reactions of organic xenobiotics. Journal of cheminformatics, 8(1), 68.

[25] Hu, L. L., Chen, C., Huang, T., Cai, Y. D., & Chou, K. C. (2011). Predicting biological functions of compounds based on chemical-chemical interactions. PloS one, 6(12), e29491.

[26] Sharma, A. K., Jaiswal, S. K., Chaudhary, N., & Sharma, V. K. (2017). A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota. Scientific reports, 7(1), 9751.

[27] Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). Biochemistry, ; W. H.

[28] Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). Lehninger principles of biochemistry. Macmillan.

[29] Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., & West, M. (2007). Generative or discriminative? getting the best of both worlds. Bayesian statistics, 8(3), 3-24.

[30] Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. Foundations and Trends® in Machine Learning, 4(4), 267-373.

[31] CRF++: Yet another CRF toolkit (2005)

[32] Sokolovska, N., Lavergne, T., Cappé, O., & Yvon, F. (2010). Efficient learning of sparse conditional random fields for supervised sequence labeling. IEEE Journal of Selected Topics in Signal Processing, 4(6), 953-964.

[33] Förster, J., Famili, I., Fu, P., Palsson, B. Ø., & Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome research, 13(2), 244-253.

[34] N. Okazaki. CRFsuite: A fast implementation of conditional random fields (CRFs), 2007.

Milana Grbić received her MSc degree (2016) in Mathematics at the Faculty of Mathematics at the University of Belgrade. She is a PhD student in the field of data mining in bioinformatics. Her research interests include data classification, data mining and bioinformatics.

‘