Bioinformatics Data Mining Tool Using Data Collected from Red Blood Cells Hemolysate

Mahmoud Rafea
Central Lab of Agriculture Expert Systems (CLAES)
Giza, Egypt
mahmoud@claes.sci.eg

Heba Zaki
Central Lab of Agriculture Expert Systems (CLAES)
Giza, Egypt
heba_zaki@claes.sci.eg

Torky Sultan
Faculty of Computers and Information, Helwan University
Cairo, Egypt
torkyibrahimsultan@hotmail.com

Abstract — **The mathematical model described in this paper is based on a discovery of a phenomenon related to red blood cells. In this phenomenon, the hemolysate of red blood cells reacts with antibodies from the plasma of the same patient. Using proteomics approach to identify those hemolysate antigens and then build a database containing those antigens can help in diagnosis, prognosis, and treatment of disease disorders. In this paper, algorithms and a tool, based on the mathematical model and the database, are described. The tool is tested using hypothetically generated data and it achieved satisfying results as it detected the proposed diseases.**

Keywords- Red Blood Cells; Antigens; Proteomics; Data Mining Tool; Bioinformatics

I. INTRODUCTION

Recently, a phenomenon related to protein content of Red Blood Cell (RBC) has been observed [1]. It was noticed that the plasma contains antibodies against some of RBC proteins, which are contained within RBC of the same person. By selecting patients suffering from open TB, it was found that some proteins of mycobacterium tuberculosis exist inside RBC and react with the patient plasma. (Under publication)

The idea behind the paper is to build a database which will be extremely huge (100 million records or so) from patients that have been investigated and diagnosed and quit good number of patients that proved to be normal, e.g., patient coming to hospitals in accidents. The patient record consists of diagnosis and a set of proteins that could be separated and identified from hemolysate using affinity chromatography and the patient own plasma antibodies as a ligand. This data can be mined for very valuable knowledge.

Building this database will take time. We need to have some data to help in testing the database and the data mining tool. Hypothetical data for patients' records can be generated to test the database.

Consequently, identifying proteins of RBC that reacts with self-antibodies and storing the identity of those proteins in a database for different disease disorders and normal individuals will help in many directions.

In this paper we will assume hypothetical disease conditions with hypothetical proteins. In effect, a mathematical model that describes the phenomena is needed to both generate the hypothetical data and build the clinical computer tool that will exploit the collected data.

The next section contains a background about laboratory methods used to identify proteins of RBC. The mathematical model is described in section III. In section IV, the algorithms of the tool are described. In section V, the scenario of the tool in clinical laboratory is described.

II. LABORATORY METHODS USED TO IDENTIFY PROTEINS OF RBC

Proteomics is a huge and complex field. The main mission of proteomics science is to identify and characterize protein expression in biological systems. It encompasses very powerful technologies and platforms, such as chromatography, electrophoresis and mass spectrometry [2].

The patient sample is collected on anticoagulant. RBC and plasma are separated. The plasma IgG is separated and then used as ligand in immunoaffinity chromatography to separate hemolysate antigens. The collected antigens are separated by two dimension electrophoresis and identified by mass spectrometry.

A. Immunoaffinity chromatography

Immunoaffinity chromatography utilizes antigens or antibodies as ligands (sometimes referred to as adsorbents, immunoadsorbents or immunosorbents) to create highly selective media for affinity purification. Antibodies are extremely useful as ligands for antigen purification, especially when the substance to be purified has no other apparent complementary ligand. Similarly, highly purified antigens or anti-antibodies can provide highly specific ligands for antibody purification.

Immunoaffinity media are created by coupling the ligand (a pure antigen, an antibody or an anti-antibody) to a suitable matrix. The simplest coupling is via the primary amine group of the ligand, using NHS-activated Sepharose or CNBr-activated Sepharose [3].

B. Two Dimension Electrophoresis

2D gel electrophoresis is a high-resolution technique for decomposing protein complexes of tenths of polypeptides. Proteins are separated according to both isoelectric point (pI) and molecular mass (Mr), by a combination of isoelectric focusing and electrophoresis respectively [4]. Spots are detected using color stains, bright coloring or radioactive labels. Electrophoresis becomes the most commonly used bidimensional protein separation method in proteomics.

978-1-4244-8843-8/10/ $ 26.00 © 2010 IEEE

C. Mass Spectrometry

Mass spectrometry has wide use in the field for protein identification and profiling experiments. The protein identification is based on detection and measurement of ionized peptides. Peptides are measured as a function of mass-to-charge ratios, termed m/z values. Peaks can be identified when measured intensities are plotted against m/z values. The location of a peak reflects the chemical composition of the protein and this can be used for protein identification, and the amplitude of a peak carries information on the protein existence, which is the basis of profiling [5].

III. Mathematical Model

It consists of four main parts; definitions of symbols, model of diseases caused by microorganisms, tumors, or foreign proteins, model of diseases caused as a result of missed tissue proteins, and model of diseases of unknown cause (Idiopathic).

A. Definitions

Let the assumption of this work be as the following:

p_i: protein amino acid sequence, where i = 1 .. n
d_j: health state, i.e., normal or disease name, where j = 1 .. m
$P = \{p_1, \ldots, p_n\}$, Set of all proteins
$D = \{d_1, \ldots, d_m\}$, Set of all diseases

P_p: patient proteins where $P_p \subset P$ where p is the patient ID
O_p: (p_i, d_j), ordered pair of patient prepresented by protein sequence (i) and health state (j).

B. Model of Diseases caused by microorganisms, tumors, or foreign proteins

$P_{dj} = \cap\{P_p\}_{dj}$

Where P_{dj} is the set which contains all common proteins associated with d_j.

$P_{normal} = \cup\{P_p\}_{normal}$

Where P_{normal} is the set which contains proteins associated with normal.

P'_{normal} such that $\forall$ p in P_{normal} if the number of occurrence of $p \in P_{normal}$ is less than 5% of the total number of p in P_{normal} then remove p from P_{normal}.

$P'_{dj} = P_{dj} - P'_{normal}$

Where P'_{dj} is the set which contains proteins that can be used as biomarker or vaccines.

C. Model of Diseases caused as a result of missed tissue proteins

$P^u_{dj} = \cup\{P_p\}_{dj}$

The result is the set which contains all proteins associated with d_j

$P''_{dj} = P_{normal} - P^u_{dj}$

The result is the set which contains proteins that can be used:

- To detect circulating auto-antibodies.
- To treat case with the proteins that give positive reaction.

D. Model of Diseases of unknown cause (Idiopathic)

- There are many diseases that are identified as idiopathic.
- Those diseases can be caused due to existence of abnormal protein or absence of tissue proteins.
- Applying data mining methods (III.*B* and III.*C*) can help to identify new diseases and treat patients appropriately.

IV. The Tool Algorithms

In this section, the algorithms that implement the previously mentioned model are explained. They are divided into three stages; data composition, Biomarkers detection, and autoantibodies detection algorithms.

A. Data Composition Algorithm

An artificial composition of proteomics data using amino acids is described here through an algorithm that produce dependable step by step generated data: Proteins, Abnormal Proteins, Normal cases, Diseases, and Patients. All the formulated data are stored in Database Tables.

- *Proteins*: from the twenty standard amino acids, 1000 different proteins are generated; each protein is assumed to have length varying among 50 and 300 amino acids that are randomly chosen.
- *Abnormal Proteins:* Depending on the previous created 1000 proteins, about 5%(50) proteins of them are chosen randomly to represent the set of Abnormal Proteins. And so, the remaining 95%(950) proteins represent the set of Normal Proteins.
- *Normal Cases:* Normal cases from the 950 Normal Proteins are composed given that the length of each normal case is varying among 15 –25 NON REPEATABLE normal Proteins and this length is determined randomly for each normal case. The normal cases are built depending on each other (with random similarity Percent from 5% to 40% between them).

Example: Normal Case (NC)

NC1: select random length (15 –25) ➔ 15 Proteins
    Pick 15 Proteins randomly from the 950 normal Proteins
NC2: Select random percentage (5% - 40%) of length of previous NC1
    Ex: 5% of 15 Proteins ➔ 1 Protein
    Pick 1 Protein randomly from NC1 to be embedded into NC2
    Select random length of NC2 ➔ 20 Proteins
    Pick the remaining 19 Proteins (20 - 1) randomly from the 950 normal Proteins
NC3: Make UNION between NC1 & NC2
    Pick random percentage from UNION to be embedded into NC3
    Select random length of NC3 ➔ 25 Proteins
    Pick the remaining Proteins needed for NC3 randomly from the 950 normal Proteins.

Stop Condition:
The length of the UNION List is 95% (950) Proteins so all the normal proteins are used and covered.
Important condition:
If the length of randomly selected similarity percentage of the UNION List is bigger than the (1/2) length of the currently composed NC then only 50% of this percentage chosen from UNION List is picked.
Ex: IF NC Len=17, Sim Percent 25% = 20 P THEN Take only 20/2 = 10 P.

- *Diseases:* Continuing our work by composing DISEASES cases from the 50 Abnormal Proteins previously composed given that the length of each disease is varying among 1–5 NON REPEATABLE Abnormal Proteins and length is determined randomly.

Example: Disease Case (DC)
DC1: Select random length (1-5) ➔ 2 Proteins
Pick 2 Proteins randomly from the 50 abnormal Proteins
DC2: Select random length of DC2 ➔ 3 Proteins.
Pick 3 Proteins randomly from the 50 abnormal Proteins, and verify uniqueness among the composed Diseases.
Make UNION between DC1 & DC2
DC3: Select random length of DC3 ➔ 1 Protein.
Pick 1 Protein randomly from the 50 abnormal Proteins, and verify uniqueness.
Make new UNION between old UNION & DC3
Stop Condition:
The length of the UNION List is (50) Proteins so all the abnormal proteins are used and covered.

- *Patients:*

- Building patients records is the last step in the first stage of data composition algorithm; it depends on mixing the normal cases with diseases.
- The selection of a normal case for modification is based on a Boolean random (0-1) to reflect resistance to infection.
- Then the normal case is selected, the disease protein sequence is either inserted or replaces a sequence, and this is achieved randomly.
- At each patient creation, a normal case is selected randomly then we decide to use it in the patient composition or not by a random Boolean. IF 0 then leave it but if 1 take it.
- Also for each patient case we will choose randomly a disease from the composed diseases to represent infection of the patient being formed
- The selected Normal Case is converted to Patient case by two ways: 1- insertion (Addition) of the selected Disease protein sequence into the selected Normal Case protein sequence OR 2- replacing part of the selected Normal Case protein sequence by the selected Disease protein sequence, and the location of replacement is chosen randomly.

Stop Condition:
All the Diseases and Normal Cases are used and covered.

B. Biomarkers Detection Algorithm

Biomarkers or vaccines of Diseases caused by microorganisms, tumors, or foreign proteins can be detected using data mining classification algorithm pseudo code that depends on the previous stage algorithm by using the composed ***Normal Cases*** and ***Patients*** through four dependable steps that are described below:

- *Step (1): Detecting P_{dj}*

Its mission is to find for each disease (d_j) the set of all common shared proteins in the patients' records suffering from this disease. "Fig. 1" shows algorithm pseudo code of this step.

```
Set DiseasesLst the list of all Diseases
Initialize CommonProteins empty lists
 with length of DiseasesLst
Initialize AllProteins empty list
loop  from 1 to DiseasesLst count
 Set CommonProteins[Cur_Disease]empty list
 Set Patient_DisList the list of patients
   associated with Cur_Disease
 Set AllProteins the list of all proteins
   in Patient_DisList
 loop from 1 to AllProteins count
  if(Cur_Protein Occurance eual_to
     Patient_DisListcount)
    Add Cur_Protein to
      CommonProteins[Cur_Disease]
   End if
  End loop
 Store CommonProteins[Cur_Disease]
End loop
```

Figure 1. *P_{dj} Detection Algorithm*

- *Step (2): Collecting P_{normal}*

Its mission is to retrieve all the proteins sharing into the normal cases records. "Fig. 2" shows algorithm pseudo code of this step.

```
Set Normal_Variants the list of all Normal
 Variants
Initialize Collected_NProteins empty list
loop  from 1 to Normal_Variants count
  Set ProteinList list of proteins exist
  in
   Cur_Normal_Variant
  Add proteins in ProteinList to
   Collected_Nproteins
End loop
Store Collected_Nproteins
```

Figure 2. *P_{normal} Collection Algorithm*

- *Step (3): Collecting P'_{normal}*

It is a subset of the P normal list produced in the previous step. It excludes the proteins from P normal that have sharing occurrence less than 5% of the normal cases. The remained proteins are considered pure normal proteins, which are (P'_{normal}). "Fig. 3" shows algorithm pseudo code of this step.

```
Set PNormal_List the list of all P Normals
Set Normals_List the list of all Normal
  Variants
Initialize PDash_List empty list
loop  from 1 to Pnormal_List count
  Set count to 0
  loop  from 1 to Normals_List count
    if (Cur_Normal contains Cur_Pnormal)
      Set count to count+1
    End if
  End loop
  Set Percent to(count/Normals_List
    count)*100
  if (Percent greater_than_OR_equal to 5)
    Add Cur_Pnormal to PDash_List
  End if
End loop
Store PDash_List
```

Figure 3. P'_{normal} *Collection Algorithm*

- *Step (4): Discovering P'_{dj}*

This is the last step in biomarkers detection stage and it depends on Steps (1) and (3). It excludes the set of (P'_{normal}) proteins exists in (P_{dj}) proteins for each disease (d_j). The result is for each disease (d_j) a set of proteins that can be used as biomarkers or vaccines for this disease. "Fig. 4" shows algorithm pseudo code of this step.

```
Set PDash_List the list of all P Dash
  Normal proteins
Set Dis_List the list of all exist Diseases
  into PIntersection_Dj
Initialize BiomarkersList empty lists with
  length of Dis_List
Initialize PIntersection_dj_List empty list
loop  from 1 to Dis_List count
  Set PIntersection_dj_List list of
   intersection proteins of Cur_Disease
  loop  from 1 to PIntersection_dj_List
   count
    if(PDash_List not_contain
       Cur_PIntersection_dj)
      Add Cur_PIntersection_dj to
       BiomarkersList[Cur_Disease]
    End if
  End loop
  Store BiomarkersList[Cur_Disease]
End loop
```

Figure 4. *P'_{dj} Discovery Algorithm*

C. *Auto-antibodies Detection Algorithm*

Antibodies or proteins give positive reaction of diseases caused as a result of missed tissue proteins can be detected using data mining classification algorithm pseudo code that depends on the previous stage algorithms by using the resulting (P_{dj}) and (P_{normal}) through two dependable steps that are described below:

- *Step (1): Detecting P^{u}_{dj}*

Its mission is to find for each disease (d_j) the set of all proteins in the patients' records suffering from this disease. "Fig. 5" shows algorithm pseudo code of this step.

```
Set DiseasesLst the list of all Diseases
Initialize CommonProteins empty lists with
  length of DiseasesLst
loop  from 1 to DiseasesLst count
  Set CommonProteins[Cur_Disease] empty
   List
  Set Patient_DisList the list of patients
   associated with Cur_Disease
  Set CommonProteins[Cur_Disease] the list
   of all proteins in Patient_DisList
  Store CommonProteins[Cur_Disease]
End loop
```

Figure 5. *P^{u}_{dj} Detection Algorithm*

- *Step (2): Discovering P''_{dj}*

It is the last and result step in this stage, and it depends on the previous step (P^{u}_{dj}) and Step (2) in Stage (1) (P_{normal}). This step gets all the missed normal proteins for each disease (d_j) by excluding all the proteins participating into patients' records of this disease (P^{u}_{dj}) from the set of all proteins participating into normal cases records (P_{normal}). The result is: a set of proteins that can be used to detect circulating auto antibodies or to treat case with the proteins that give positive reaction for each disease (d_j). "Fig. 6" shows algorithm pseudo code of this step.

```
Set PNormal_List the list of all P Normals
Set DiseasesLst the list of all Diseases
Initialize P2Dash_dj_List empty lists with
  length of DiseasesLst
Initialize PUnion_dj empty list
loop  from 1 to DiseasesLst count
  Set P2Dash_dj_List[Cur_Disease] empty list
  Set PUnion_dj the list of Punion proteins
   of Cur_Disease
  loop from 1 to PNormal_List count
    if (PUnion_dj not_contains Cur_PNormal)
      Add Cur_Pnormal to
         P2Dash_dj_List[Cur_Disease]
    End if
  End loop
  Store P2Dash_dj_List[Cur_Disease]
End loop
```

Figure 6. *P''_{dj} Discovery Algorithm*

V. SCENARIO OF THE TOOL IN CLINICAL LABORATORY

Patients' blood samples will be collected on anticoagulant. RBC and plasma are then separated in different tubes. Plasma is used as a ligand in immunoaffinity chromatography to separate hemolysate antigens that can bind to plasma antibodies. The separated antigens are identified by MS and stored in the database indexed by the patient disorder.

In the same time, queries are done to verify the diagnosis and get a prognosis and a recommended treatment component. The following formulas describe the usage of this model in clinical practice.

- Let Dp' is the set of all discovered P'_{dj}
- Let Dp" is the set of all discovered P''_{dj}

$\forall P'_{dj} \in Dp'$, if $P'_{dj} \subset P$, then patient is diagnosed to have d_j

Else

$\forall P''_{dj} \in Dp''$, if $P''_{dj} \not\subset P$, then patient is diagnosed to have d_j

VI. CONCLUSION

RBCs contain antigens, which are related to patients' health state. In this paper, a mathematical model for exploiting this newly discovered RBC phenomenon is described. It is based on building a database of patients' diagnosis and the set of proteins identified through affinity chromatography, 2D electrophoresis, and mass spectrometry. The presented model shows the data mining approach that can be used to identify biomarkers, help in identifying proteins that can be used in treatment of disorders.

The main purpose of this tool is to help in diagnosing disease conditions that are difficult to diagnose, for instance: autoimmune disorders, chronic rejection, and idiopathic diseases. Malignancy can be very early diagnosed through direct approach using tumor specific antibodies against tumor-hemolysate antigens.

The presented application can be used as a tool in clinical laboratories to diagnose disease disorders and propose personalized treatment. The paper describes, also, the implementation algorithms of this model using pseudo code. In effect, a tool has been built and verified by generating artificial data.

REFERENCES

[1] Rafea M., M.El-Ayouby S., El Ansary M., Helmy N., Hosni J., and Gamil J. "The Security Hole of the Immune System," In the proceedings of The Middle east and North Africa (MENA) region animal wealth researches conference & the international exhibition for animal production and dairy industry animal tech expo, Cairo, Egypt, Oct. 2008.

[2] Michael R. Barnes, GlaxoSmithKline Pharmaceuticals and Ian C. Gray. "Bioinformatics for Geneticists," Copyright ©2003 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, England, West Sussex PO19 8SQ.

[3] "Affinity Chromatography Principles and Methods," Handbooks from GE Healthcare, 2007.

[4] O'Farrell, P. H. "High resolution two-dimensional electrophoresis of proteins," J. Biol. Chem. 250, 4007–4021 (1975)

[5] Hui-Huang Hsu. "Advanced data mining technologies in bioinformatics," Copyright © 2006 by Idea Group Inc.

[1] The development of such database is protected by PCT International Publication Number: WO 2009/066131 A1